UNIVERSITAT AUTÒNOMA DE BARCELONA

Anàlisi del Rendiment Acadèmic del Grau d'Estadística Aplicada

Treball Final de Grau

David Rivero Martí

Tutora: Alejandra Cabaña Nigro

Grau d'Estadística Aplicada

Juny 2017

1. Resum

Resum

L'estudi analitza el rendiment acadèmic del grau d'Estadística Aplicada de la Universitat Autònoma de Barcelona durant els anys 2010-2015 mitjançant algunes de les tècniques estadístiques que s'han treballat durant el mateix grau.

Paraules clau:

- Rendiment acadèmic
- Arbres de regressió
- Clústers

Resumen

El estudio analiza el rendimiento académico del grado de Estadísica Aplicada de la Universidad Autónoma de durante los años 2010-2015 de la Universidad Autónoma de Barcelona mediante técnicas estadísticas desarrolladas durante el mismo grado.

Abstract

In this work I wanted to express some of the knowledge acquired during my Degree in Applied Statistics at Autonomous University of Barcelona during the years 2010-2015.
To explain this procedure, I have performed differents types of statiscal analysis learned in my degree.

ÍNDEX

2. Introducció

En aquest projecte es posa de manifest el rendiment que tenen els alumnes durant el grau d'Estadística Aplicada impartit per l'UAB amb dades de 2010 (any que va començar el grau com a tal i l'any 2015).

Es posaran en pràctica alguna de les tècniques estadístiques apreses durant les assignatures del grau.

L'estudi, neix d'un projecte previ desenvolupat pel SIQ (Sistema Intern de Qualitat de la UAB).

El Sistema Intern de Qualitat de la UAB és l'eina amb que es dota la Universitat per garantir la qualitat dels seus programes formatius, establint una infraestructura de funcionament i un conjunt de processos orientats a la millora contínua.

2.1 Objectius

L'objectiu principal del treball es posar en pràctica les tècniques i coneixements adquirits durant el grau d'Estadística Aplicada.

Els objectius intrínsecs del treball són, entre altres:

- Observar quin és el rendiment acadèmic mitjà segons el curs acadèmic.
- Saber si hi ha diferències entre el rendiment del primer semestre i el segon.
- Identificar si ha millorat el rendiment del grau des de que va començar (2010) fins l'últim any del que tenim dades (2015).
- Esbrinar mitjançant anàlisi de clústers quines assignatures tenen un comportament estadísticament semblant per veure si hi ha coherència entre la distribució de les assignatures per cada curs del grau.
- Realitzar arbres de regresió per observar quines variables són les que millor expliquen el rendiment.

2.2 Les dades

Les dades necessàries per a dur a terme el aquest treball s'han extret de la Web de SIQ (Sistema Intern de Qualitat de la UAB), un cop dins del grau d'Estadística Aplicada, a l'apartat de resultats acadèmics ens trobem amb les dades que van des de l'any 2010 (quan va començar el grau) fins l'any 2015.

Val a dir que s'ha intentat aconseguir dades d'anys anteriors, és a dir, de quan la carrera era una diplomatura, però ha estat impossible i al final s'ha desestimat l'opció.

Les dades del treball consten doncs de 11 variables:

- **Codi**: Codi numèric de l'assignatura. 6 dígits.
- **Assignatura**: Nom de l'assignatura.
- **Matriculats**: Nº de matriculats.
- **Mh**: Nº de Matrícules d'Honor.
- **Exc**: Nº d'Excel.lents.
- **Nt**: Nº de Notables.
- **Ap**: Nº d'Aprovats.
- **S**: Nº de Suspesos
- **Np**: Nº de No Presentats.
- **Rend**: Aquesta variable és una de les principals a tractar en aquest treball, Cal remarcar que aquesta variable és el resultat de dividir el Nº de Superats / Nº de Matriculats.
- **Èxit**: Aquesta variable també és el resultat d'una operació entre variables.
    - En concret , la divisió del Nº de superats / Nº de presentats.
    - Al obviar els no presentats, la variable èxit tindrà uns valors sempre superiors a la variable Rendiment.
- **%np**: Tant per cent de No presentats.

<u>Variables Afegides per a l'estudi</u>

A part de les 11 variables proporcionades pel SIQ, he incorporat algunes més que m'han semblat rellevants a l'hora de fer l'anàlisi, aquestes són:

- **Any**: Figura al desplegable i anirà del 2010 al 2015.
- **Curs**: fent referència al curs acadèmic per tant prendrà els valors d'1 a 4.
- **Semestre**: fent referència al semestre al qual pertany l'assignatura (1r o 2n).
- **Presentats**: Resultat de la resta: Matriculats – No presentats.
- **Superats**: Suma de Mh+Exc+Nt+Ap.

2.3 Disseny Experimental

Per analitzar les dades del projecte, s'han fet servir diferents anàlisis.

Anàlisi descriptiva bivariant

Aquí veurem com es comporta la nostra variable resposta (Rendiment) en funció de les demés variables rellevants de l'estudi com:

- Curs
- Semestre
- Any
- Assignatures

Al final d'aquest apartat s'ha realitzat un anàlisi descriptiu d'aquelles assignatures amb el rendiment més baix del grau. És a dir, aquelles assignatures en les que més del 50% dels matriculats no arriben a superar-la.

Anàlisi de clústers

L'objectiu de l'anàlisi de clústers serà veure quines assignatures tenen un comportament semblant en funció del seu rendiment i altres variables.

Al final del capítol es presenta l'algorisme *k-means* i s'aplica a les dades el treball.

Arbres de regressió

En aquest apartat, es té com a objectiu trobar variables que discriminin la nostra variable resposta mitjançant un arbre de regressió.

Al final de l'apartat és valida l'arbre de manera interna i externa.

Per últim, anomenar els softwares utilitzats per dur a terme l'anàlisi:

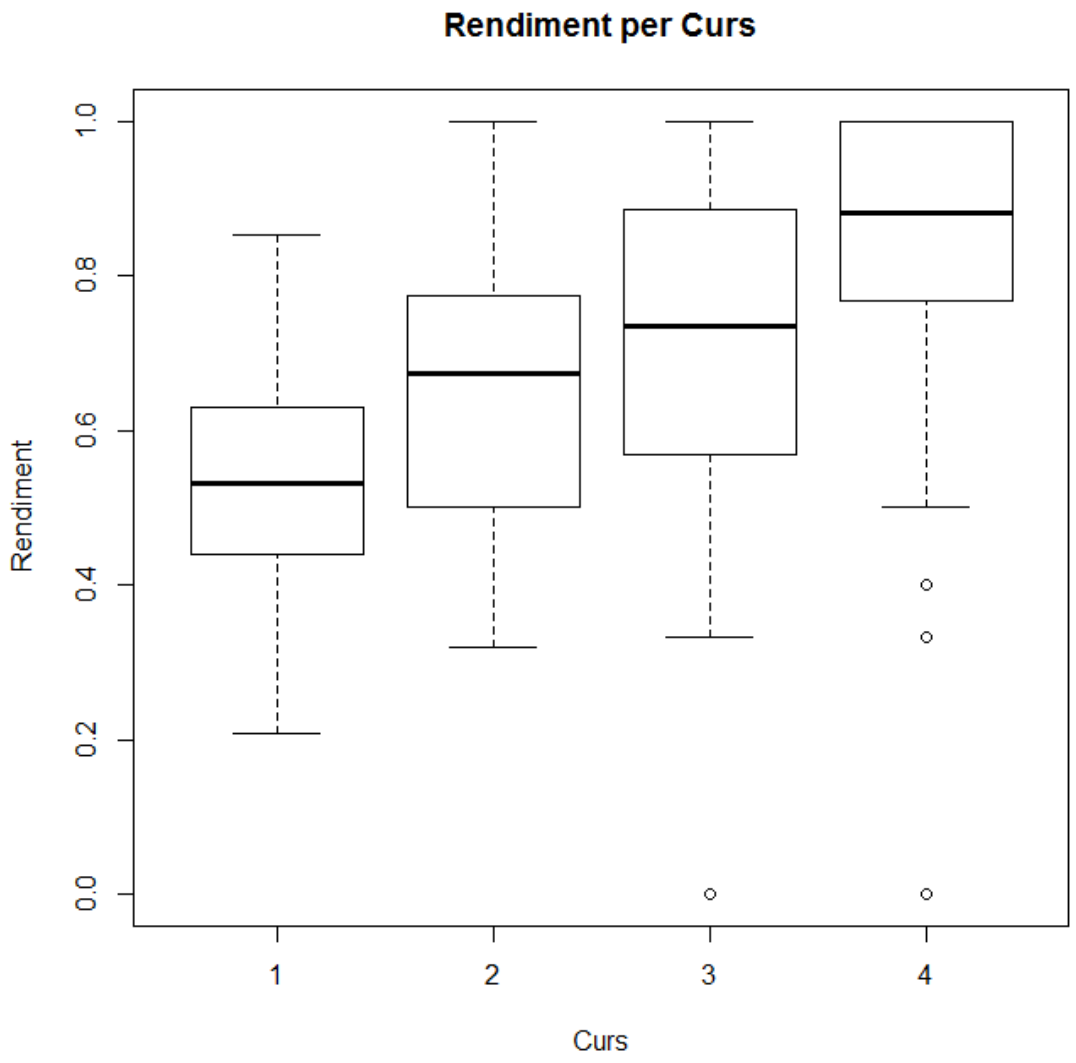
- **R**: Utilitzat per tot el treball
- **Microsoft Excel**: Per a les bases de dades i algun anàlisi bivariant.

3. Anàlisi Descriptiva Bivariant

Començarem l'anàlisi fent una breu anàlisi descriptiva bivariada sobre les variables més rellevants per al nostre estudi respecte la variable rendiment.

3.1 Rendiment per curs

FIGURA 1. BOX-PLOT DE RENDIMENT PER CURS

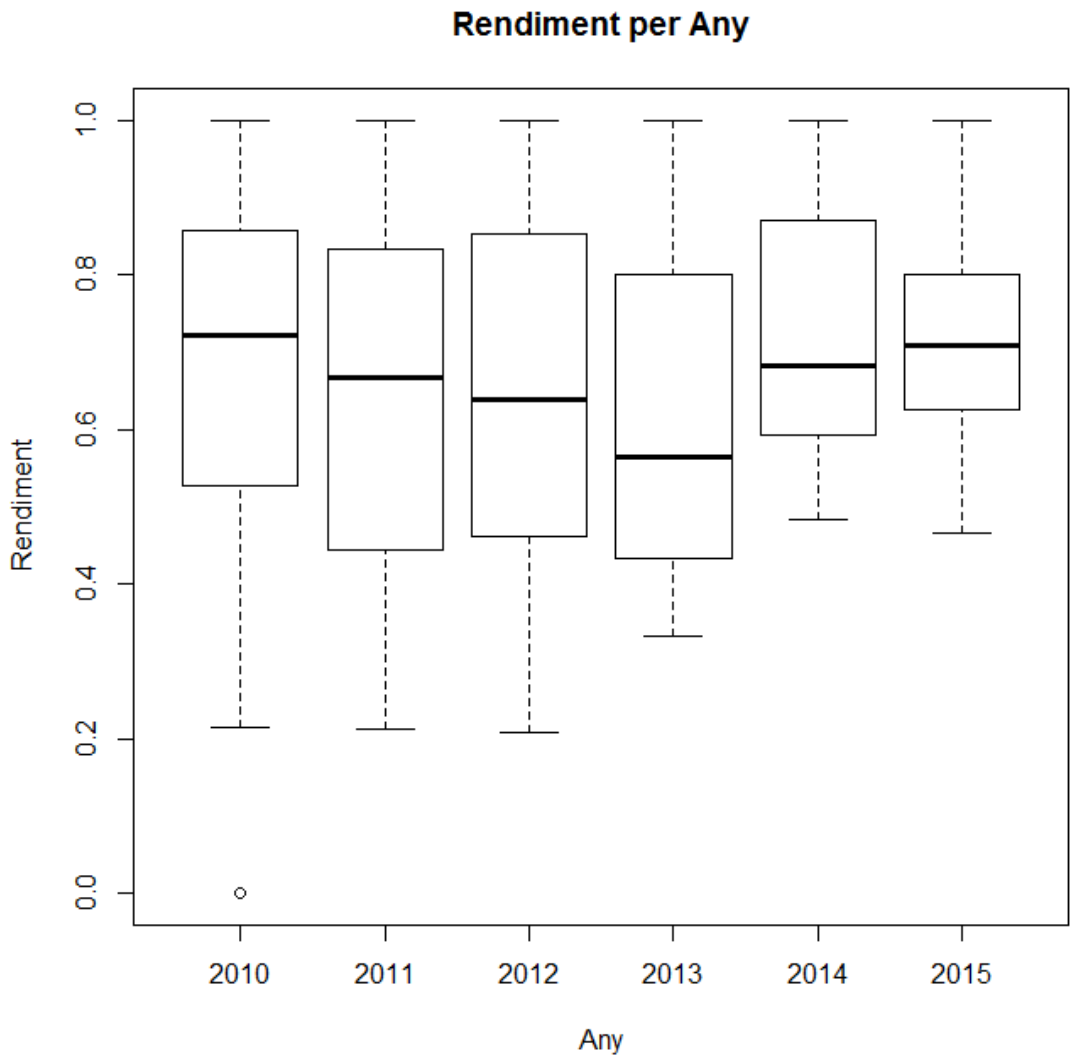


La mitjana del rendiment per al primer curs és del 53%, per al segon curs és del 65%. al tercer curs és del 70% i la més alta és pel quart curs amb un 83%.

Com podem observar al gràfic box-plot, el rendiment augmenta a mesura que passen el cursos. Als cursos 3r i 4rt observem la presència d'algun valor atípic.

3.2 Rendiment per Any

FIGURA 2. BOX-PLOT DEL RENDIMENT PER ANY



Realitzarem un model de regressió simple per interpreter amb més rigurositat l'efecte Any sobre Rendiment.

lm(formula = rend ~ Any, data = dades)

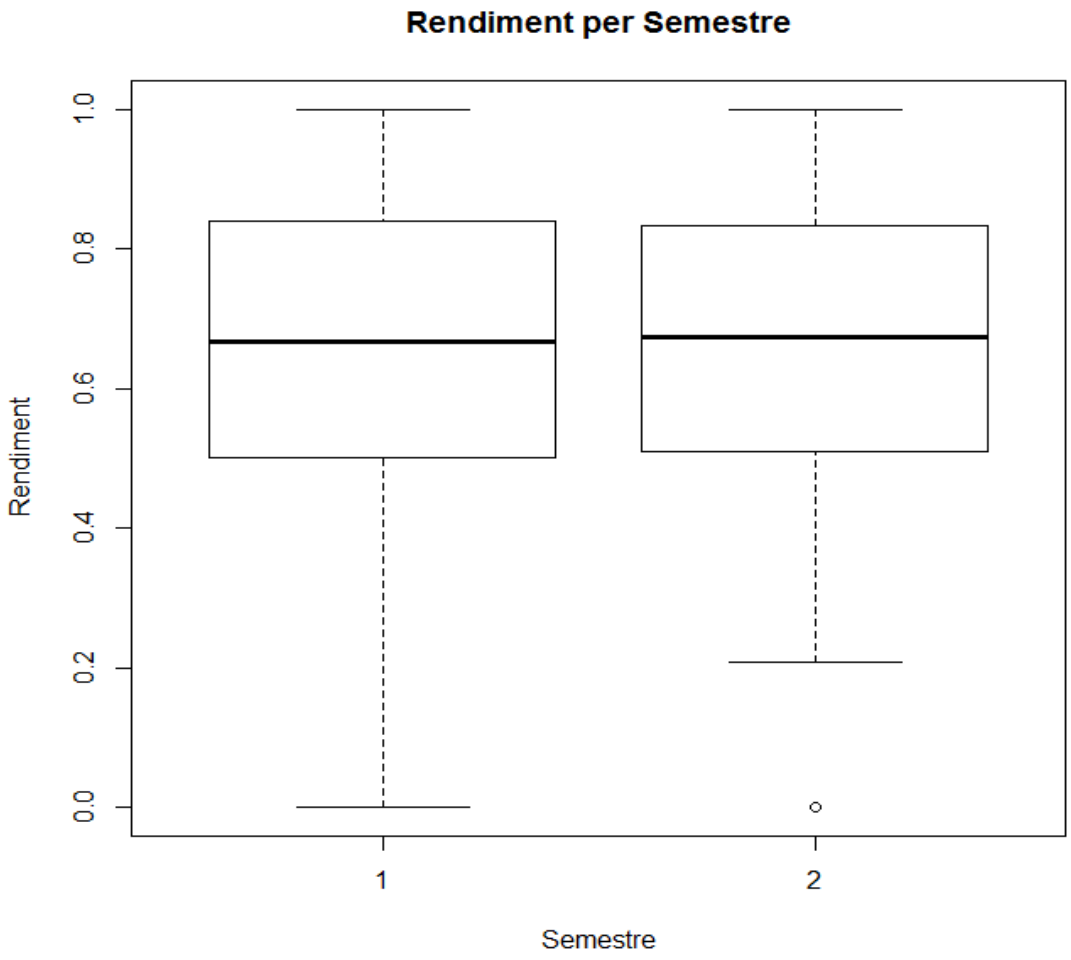
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-23.352592	17.115083	-1.364	0.174
Any	0.011936	0.008504	1.404	**0.162**

Podem observer com l'Any no és estadísticament significativa pel model amb un p-valor>0.05, per tant no es un bona variable predictora per al rendiment, amb el que podem concloure que no s'aprecien tendències anuals.

3.3 Rendiment per Semestre

FIGURA 3. BOX-PLOT DEL RENDIMENT PER SEMESTRE



Aquest gràfic és interessant ja que podem osbervar com no s'aprecien diferències notables entre el rendiment del primer i del segon semestre.

De totes maneres, procedim a realizar un t.test per assegurar-nos:

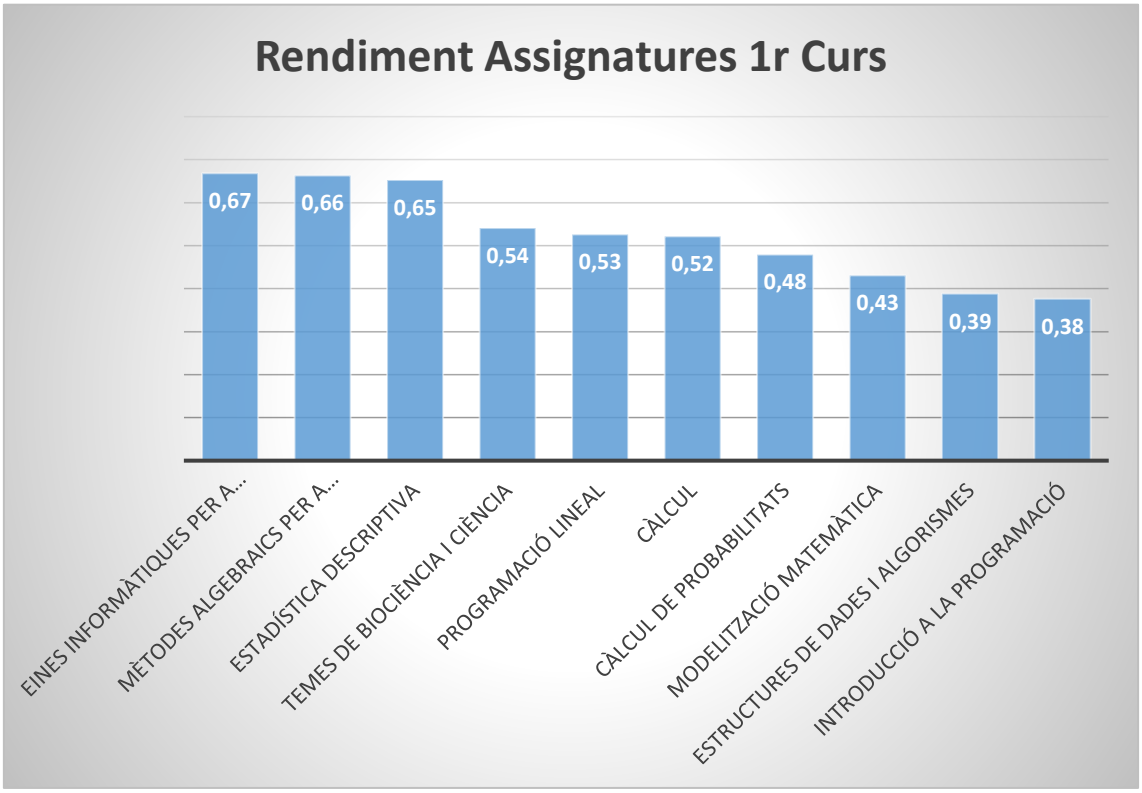
Welch Two Sample t-test
data: rend by Semestre
t = 0.030321, df = 225.32, p-value = **0.9758**
mean in group 1 mean in group 2
0.6690640 0.6681794

Efectivament, amb un p-valor > 0.05 no tenim prou evidències per rebutjar la hipótesis d'igualtat de mitjanes. Si ens fixem són pràcticament iguals.

- Mitjana del Rendiment pel 1r Semestre: 0.6690
- Mitjana del Rendiment pel 2n Semestre: 0.6681

3.4 Rendiment per Assignatures

FIGURA 4. GRÀFIC DE BARRES DEL RENDIMENT PER ASSIGNATURES DE PRIMER CURS



La mitjana del Rendiment per al primer curs és del 53%.

Les assignatures amb rendiments més alts a primer són:

- Eines Informàtiques per a l'Estadística.
- Mètodes Algebraics per a l'Estadística.
- Estadísitca descriptiva.

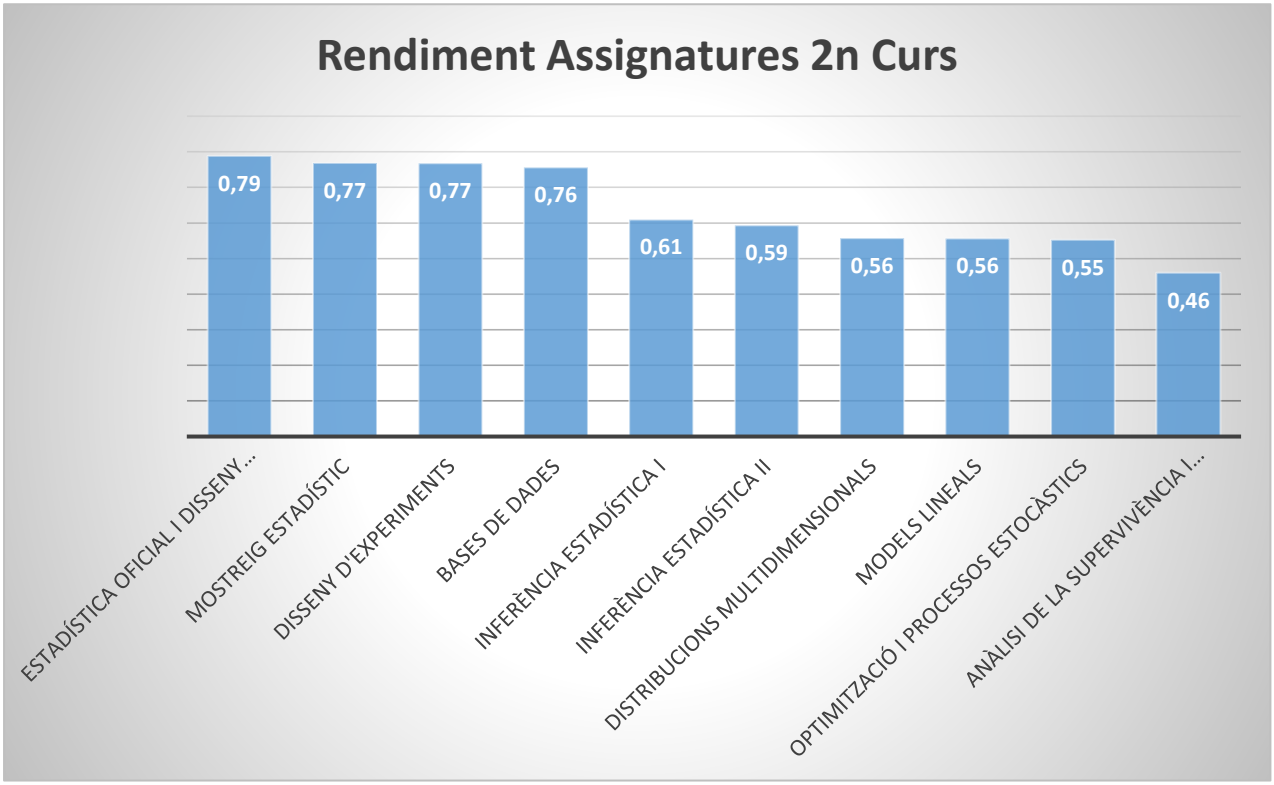
Trobem 4 assignatures amb rendiment per sota del 50%:

- Càlcul de Probabilitats.
- Modelització Matemàtica.
- Estructures de Dades I Algorismes.
- Introducció a la Programació.

Estructures de Dades i Algorismes i Introducció a la Programció són les dues assignatures amb rendiment mitjà més baix a primer curs.

Val a dir que les assignatures dues són de caire purament informàtic i per tant, estan directament relacionades entre elles.

FIGURA 5. GRÀFIC DE BARRES DEL RENDIMENT PER ASSIGNATURES DE SEGON CURS



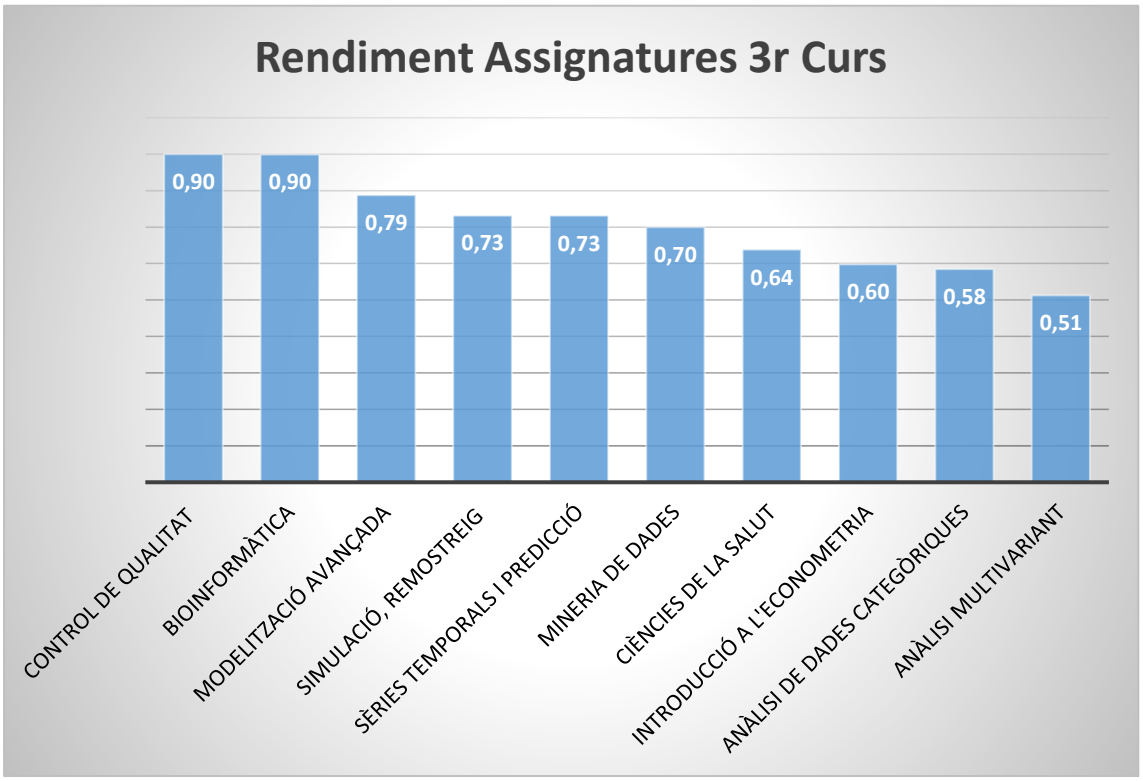
La mitjana del Rendiment per al segon curs és del 65%, millor rendiment mitjà que a primer.

Hi han 4 assignatures que es troben amb un rendiment força bo, per sobre del 70%, són:

- Estadística Oficial i Disseny d'Enquestes
- Mostreig Estadístic
- Disseny d'Experiments
- Bases de Dades

L'assignatura amb el rendiment més baix de 2n curs és la d'Anàlisi de la Supervivència i Fiabilitat, amb un rendiment mitjà per sota de 50 %, és a dir, més de la meitat dels matriculats en aquesta assignatura no la superen.

FIGURA 6. GRÀFIC DE BARRES DEL RENDIMENT PER ASSIGNATURES DE TERCER CURS



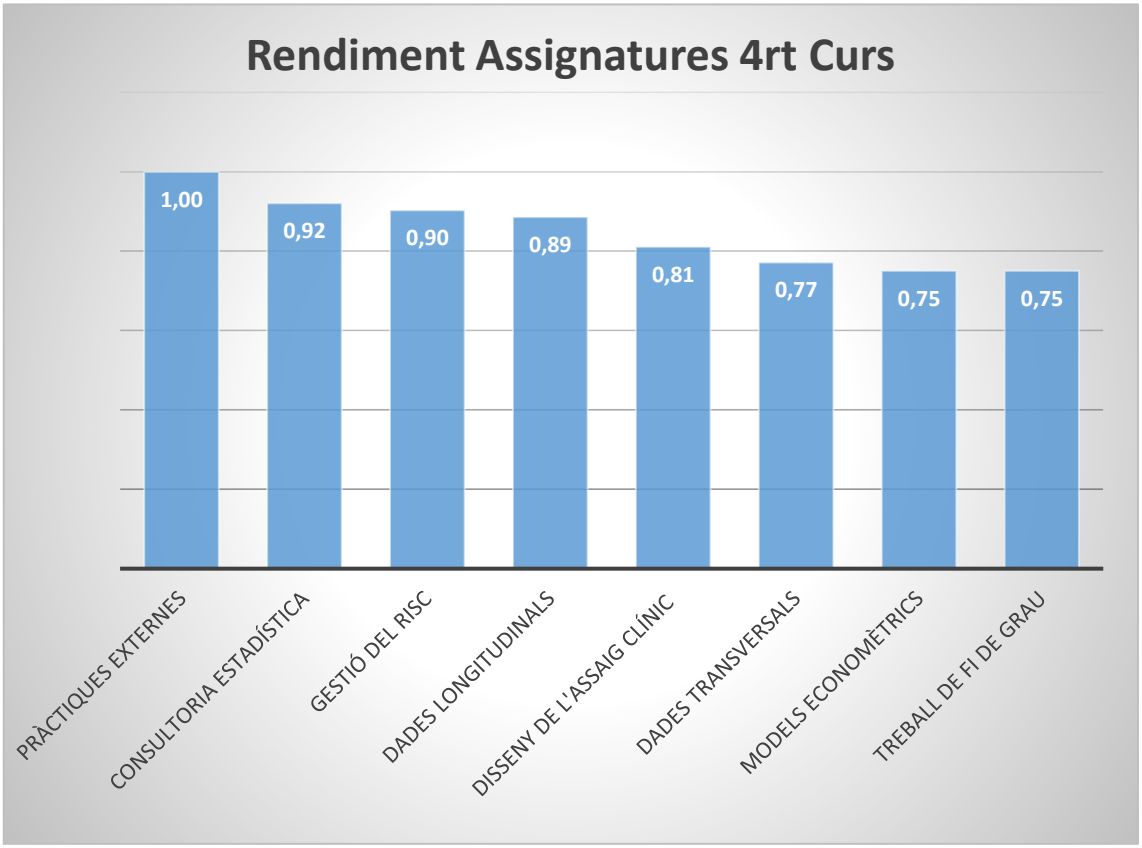
La mitjana del Rendiment per al tercer curs és del 70%, veiem doncs que el rendiment millora notablement a mesura que augmentem de curs.

A destacar 2 molts bons rendiments a les assignatures:

- Control de Qualitat i Estadística Industrial
- Aplicacions de l'Estadística a la Bioinformàtica

Per contra, l'assignatura amb el rendiment més baix de 2n curs és la d'Anàlisi Multivariant amb un rendiment del 51%. Tot i així no està per sota del 50%.

FIGURA 7. GRÀFIC DE BARRES DEL RENDIMENT PER ASSIGNATURES DE TERCER CURS



La mitjana del Rendiment per al primer curs és del 83%. Clarament el curs amb millor rendiment del grau d' Estadística Aplicada.

Val a dir però, que en aquest curs la majoria d'assignatures són optatives i que l'assignatura de Pràctiques Externes no és comparable amb altres assignatura ja que té un pes diferent, el mateix passa amb la de Treball final de Grau.

A destacar 4 assignatures amb rendiment mitjà superior al 80%:

- Consultoria Estadística
- Gestió del Risc
- Dades Longitudinals: Temes Avançats en Ciències de la Salut

3.5 Anàlisi de les assignatures amb baix rendiment

A continuació procedirem a fer un breu anàlisi descriptiu de les 5 assignatures del grau amb rendiment més baix, és a dir, per sota del 50%.

A primer curs es on trobem la majoria d'assignatures amb baix rendiment, en concret: Introducció a la Programació, Modelització Matemàtica, Càlcul de Probabilitats, Estructura de Dades i Algorismes.

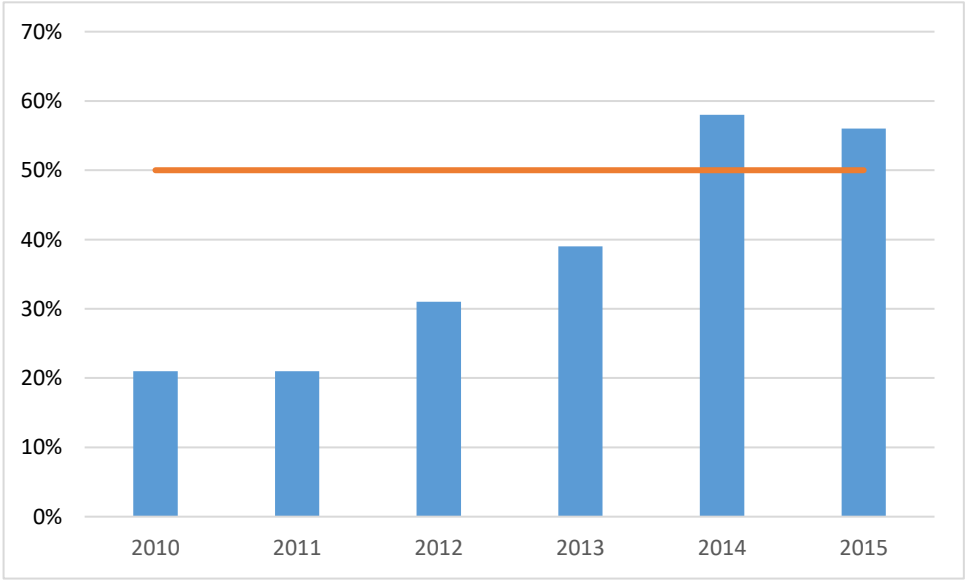
A segon curs trobem una asignatura per sota del 50% de rendiment que és la d' Anàlisi de la Supervivència i Fiabilitat.

Introducció a la Programació

TAULA 1. EVOLUCIÓ DEL RENDIMENT D'INTRODUCCIÓ A LA PROGRAMACIÓ

Any	Matriculats	mh	exc	nt	ap	s	np	%np	rend
2010	28	0	1	0	5	8	14	50%	21%
2011	47	1	1	2	6	14	23	49%	21%
2012	59	0	0	0	18	23	18	31%	31%
2013	54	2	0	4	15	13	20	37%	39%
2014	38	0	1	5	16	7	9	24%	58%
2015	45	2	1	4	18	5	15	33%	56%
Totals	**271**	**5**	**4**	**15**	**78**	**70**	**99**	**37%**	**38%**

FIGURA 8. GRÀFIC DEL RENDIMENT D'INTRODUCCIÓ A LA PROGRAMACIÓ PER ANY



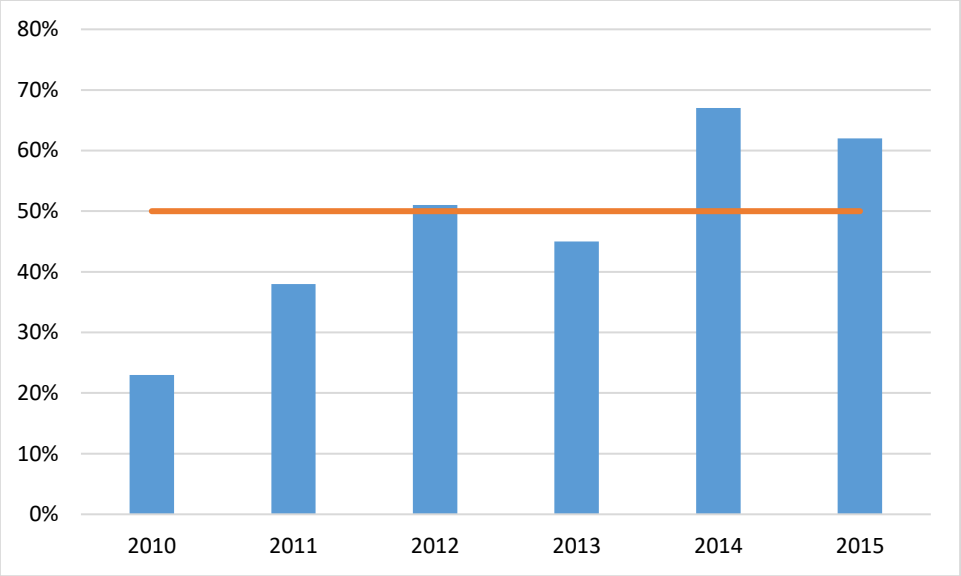
S'observa com als 4 primers anys del grau el Rendiment de l'assignatura ha estat per sota del 50% però a partir del 2014 ja superava el 50% el que ens pot fer pensar en un canvi de profesorat en l'assignatura.

Càlcul de Probabilitats

TAULA 2. EVOLUCIÓ DEL RENDIMENT DE CÀLCUL DE PROBABILITATS

Any	Matriculats	mh	exc	nt	ap	s	np	%np	rend
2010	31	0	1	3	3	10	14	45%	23%
2011	42	1	0	3	12	2	24	57%	38%
2012	51	0	1	6	19	7	18	35%	51%
2013	44	3	2	7	8	3	21	48%	45%
2014	33	1	2	11	8	1	10	30%	67%
2015	39	2	5	7	10	0	15	38%	62%
Totals	**240**	**7**	**11**	**37**	**60**	**23**	**102**	**43%**	**48%**

FIGURA 9. GRÀFIC DEL RENDIMENT DE CÀLCUL DE PROBABILITATS PER ANY



Observem uns primers dos anys amb uns rendiments molt baixos, després augmenta amb una petita devallada al 2013. El any amb millor rendiment de l'assignatura va ser el 2014 amb només un suspès.

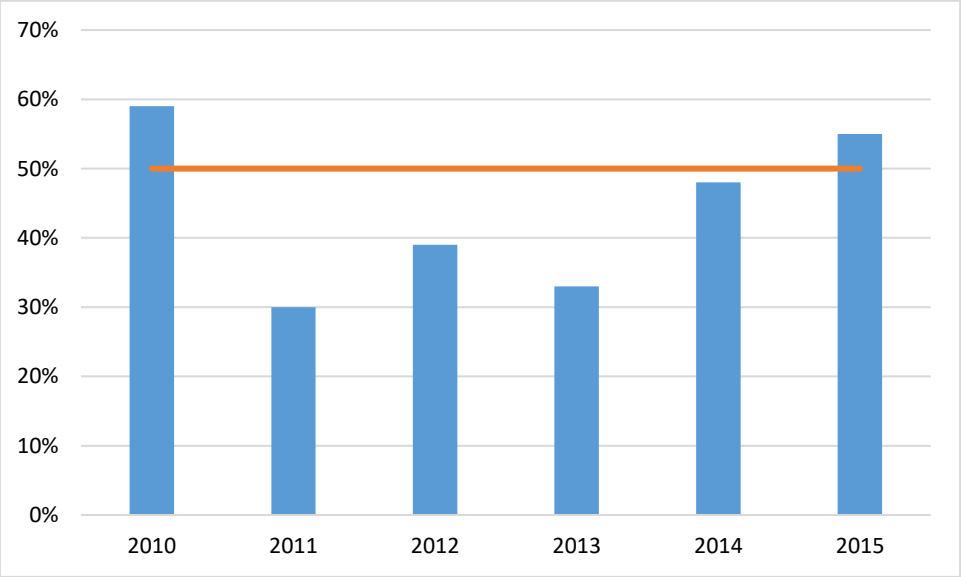
El pitjor any en quan a rendiment acadèmic va ser el 2010 amb un total de 10 suspesos i 14 no presentats

Modelització Matemàtica

TAULA 3 EVOLUCIÓ DEL RENDIMENT DE MODELITZACIÓ MATEMÀTICA

Any	Matriculats	mh	exc	nt	ap	s	np	%np	rend
2010	27	1	0	3	12	4	7	26%	59%
2011	37	0	1	1	9	6	20	54%	30%
2012	44	0	0	2	15	4	23	52%	39%
2013	42	1	0	6	7	5	23	55%	33%
2014	31	0	0	8	7	5	11	35%	48%
2015	44	4	1	7	12	7	13	30%	55%
Totals	**225**	**6**	**2**	**27**	**62**	**31**	**97**	**43%**	**43%**

FIGURA 10. GRÀFIC DEL RENDIMENT DE MODELITZACIÓ MATEMÀTICA PER ANY



Només en dos del cinc anys el rendiment ha superat el 50%, es tracta del 2010 i 2015.

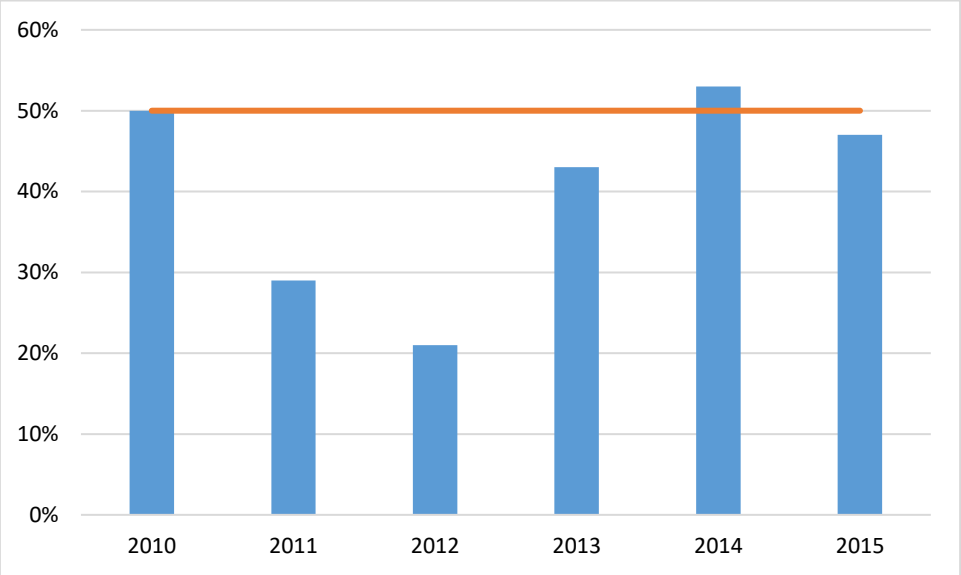
Un dels motius principals es deu a la gran quantitat de no presentats, sobretot entre els anys 2011-2013.

Estructura de Dades i Algorismes

TAULA 4 EVOLUCIÓ DEL RENDIMENT DE ESTRUCTURES DE DADES I ALGORISMES

Any	Matriculats	mh	exc	nt	ap	s	np	%np	rend
2010	24	1	0	1	10	2	10	42%	50%
2011	42	0	0	2	10	2	28	67%	29%
2012	48	0	0	3	7	11	27	56%	21%
2013	53	0	0	8	15	1	29	55%	43%
2014	30	0	0	9	7	5	9	30%	53%
2015	45	2	1	3	15	10	14	31%	47%
Totals	**242**	**3**	**1**	**26**	**64**	**31**	**117**	**48%**	**39%**

FIGURA 11 EVOLUCIÓ DEL RENDIMENT DE ESTRUCTURES DE DADES I ALGORISMES PER ANY



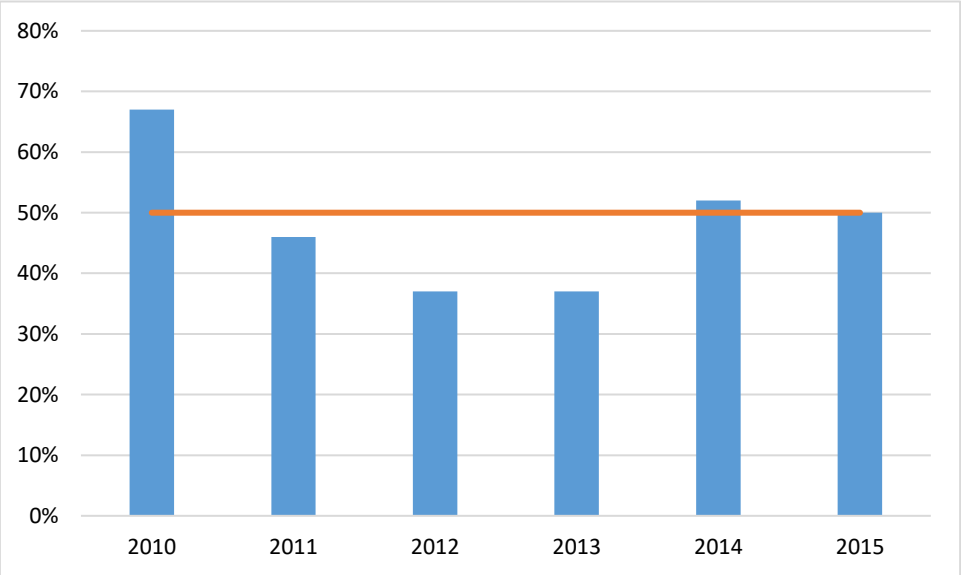
Veiem com al 2011 i 2012 va haver un rendiment molt baix, degut en part a la gran quantitat de no presentats. El 2013 va aprovar més del doble d'alumnes respecte el 2013 però un altre cop degut al gran nombre de no presentats el rendiment no va arrivar al 50%. 2014 i 2015 millora el rendiment, però s'observa com en termes generals, és una asignatura amb baix rendiment.

Anàlisi de la Supervivència i Fiabilitat

TAULA 5 EVOLUCIÓ DEL RENDIMENT D'ANÀLISI DE LA SUPERVIVÈNCIA I FIABILITAT

Any	Matriculats	mh	exc	nt	ap	s	np	%np	rend
2010	9	0	0	3	3	2	1	11%	67%
2011	28	0	0	1	12	3	12	43%	46%
2012	30	0	0	1	10	10	9	30%	37%
2013	27	0	1	1	8	7	10	37%	37%
2014	31	0	0	2	14	9	6	19%	52%
2015	38	0	0	1	18	14	5	13%	50%
Total	**163**	**0**	**1**	**9**	**65**	**45**	**43**	**26%**	**46%**

FIGURA 12 EVOLUCIÓ DEL RENDIMENT D'ANÀLISI DE LA SUPERVIVÈNCIA I FIABILITAT PER ANY



De les 5 assignatures analitzades és la que té el tant per cent de no presentats més baix, tot i així en qüasi 4 dels 6 anys, el rendiment és menor al 50%.

L'any amb millor rendiment de l'assignatura va ser al 2010 amb un 67%, val a dir que en aquest any només hi havien 9 matriculats ja que l'assignatura correspon al segon curs. I el grau com a tal va començar al 2010. És a dir aquests 9 matriculats pertanyien a l'antiga diplomatura amb lo qual no és un any comparable amb els demés.

4. Anàlisi de Clústers

4.1 Introducció

L´anàlisi de grups (cluster analysis) és un conjunt de tècniques estadístiques que serveixen per determinar grups internament homogenis, però diferents entre ells.

Aquestes tècniques classifiquen individus o objectes tenint en compte totes les variables de l´anàlisi, sense referir-se al comportament d´una variable criteri específica.

Per determinar la similitud entre objectes o individus existeixen diverses mesures d'associació però una de les més utilitzades és la distància euclidiana.

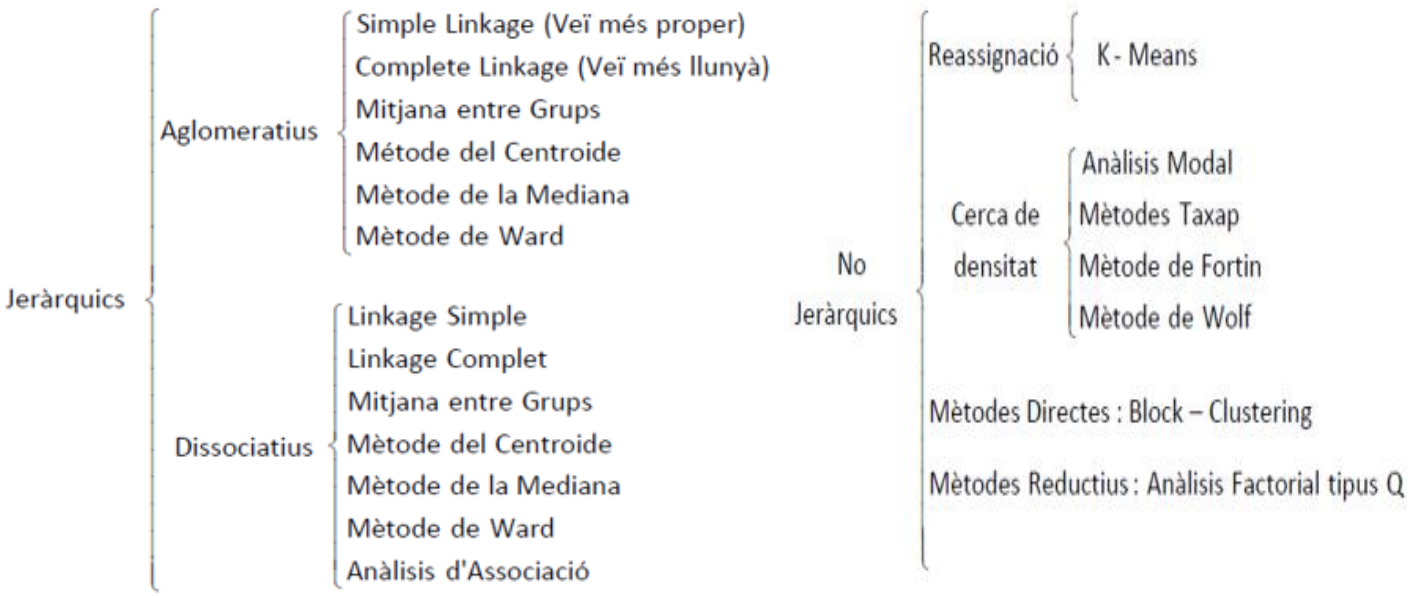
La distància euclidiana mesura la distància entre dos punts en un espai geomètric d' n dimensions. Donats dos punts i i j, en un espai n-dimensional, la distància entre ells, dij, es formula de la manera següent:

$$d_{ij} = \left[\sum_{k=1}^{n} (X_{ik} - X_{jk})^2 \right]^{1/2}$$

On *Xik* i *Xjk* són les projeccions dels punts *i* i *j* sobre la dimensió *k (k = 1, 2, 3, ... n).*

Tenim diferents mètodes per formar clústers, els classifiquem entre Jeràrquics, i no Jeràrquics. Comentarem alguns en detall més endavant.

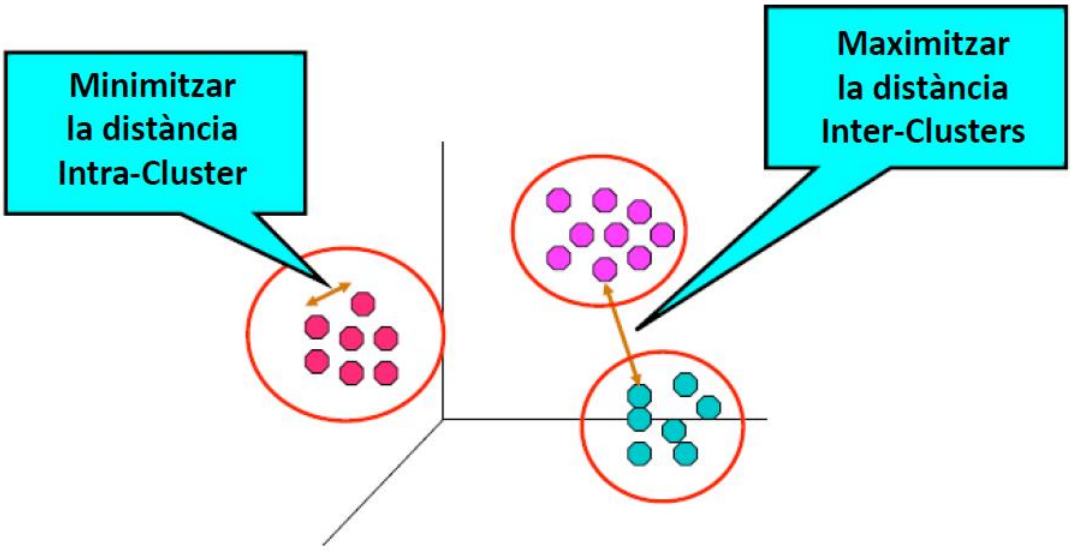
FIGURA 13. MÈTODES D'ANÀLISI CLÚSTER



4.2 Objectiu

L'objectiu principal quan fem anàlisi de clústers és obtenir homogeneïtat dins les clases i separació entre elles.

FIGURA 14. OBJECTIU DE L'ANÀLISI DE CLÚSTERS REPRESENTAT GRÀFICAMENT



4.3 Aplicació a les nostres dades

Per al nostre treball farem servir els mètodes més coneguts i utilitzats, en concret seran:

Complete, Single, Average, Ward.

Amb el software estadístic R representarem els correspontents dendogrames, que són els gràfics on es mostra el procés d'agrupament entre els casos i la distància amb la que es produeix cada agrupament.

Ordenarem que ens agrupi les assignatures en 4 grups per poder fer millor el símil amb la realitat. És a dir, veure si están distribuides segons el curs académic o no.

Escollirem el mètode que millor s'ajusti a les nostres dades i finalment farem una breu presentación del mètode no jeràrquic *K-means*.

4.3.1 Mètode complete

Aquest mètode considera que la distància o similitud entre 2 clústers, s'ha de medir segons els seus elements més dispars, és a dir, la distància ve donada per la màxima distància entre els seus components.

FIGURA 15. DENDOGRAMA PEL MÈTODE COMPLETE ASSIGNANT 4 CLUSTERS, ETIQUETEM PER ASSIGNATURA

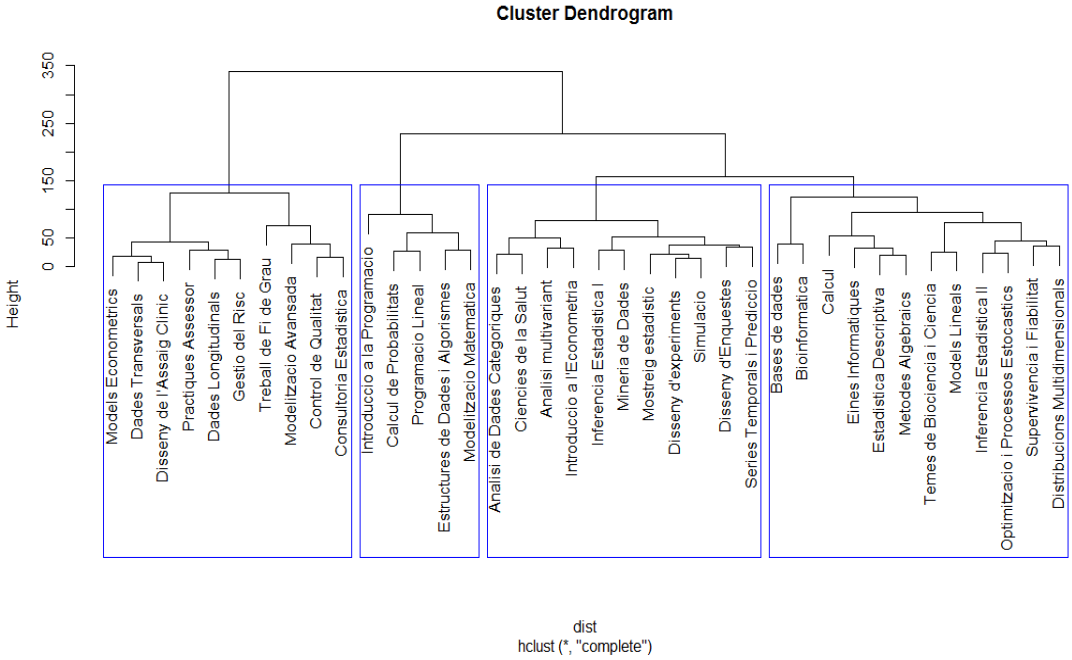
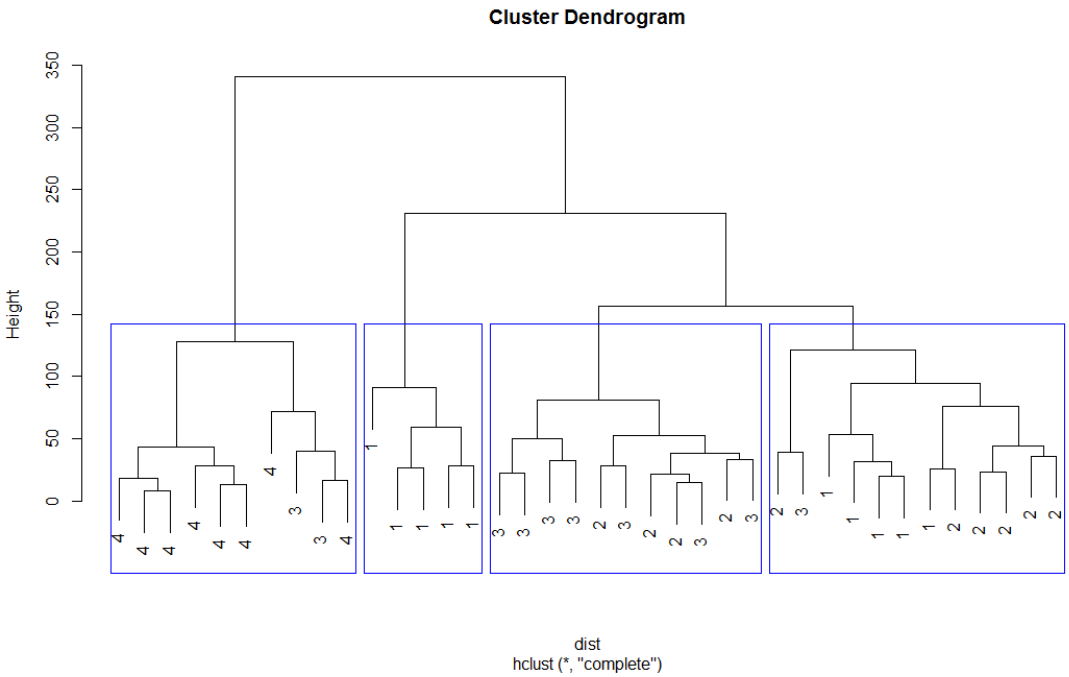


FIGURA 16. DENDOGRAMA PEL MÈTODE COMPLETE ASSIGNANT 4 CLUSTERS, ETIQUETEM PER CURS



- **Clúster 1**:
    - Models Economètrics (4rt)
    - Dades Transversals: Temes Avançats en Ciències de la Salut (4rt)
    - Disseny de l'Assaig Clínic i Metodologia Estadística Aplicada (4rt)
    - Pràctiques Externes (4rt)
    - Dades Longitudinals: Temes Avançats en Ciències de la Salut (4rt)
    - Gestió del Risc (4rt)
    - Treball de Fi de Grau (4rt)
    - Modelització Avançada (3)
    - Control de Qualitat i Estadística Industrial (3)
    - Consultoria Estadística (4rt)
- **Clúster 2**:
    - Introducció a la Programació (1r)
    - Càlcul de Probabilitats (1r)
    - Programació Lineal (1r)
    - Estructures de Dades i Algorismes (1r)
    - Modelització Matemàtica (1r)
- **Clúster 3**:
    - Anàlisi de Dades Categòriques (3r)
    - Aplicacions de l'Estadística a les Ciències de la Salut (3r)
    - Anàlisi Multivariant (3r)
    - Introducció a l'Econometria (3r)
    - Inferència Estadística I (2n)
    - Mineria de Dades (3r)
    - Mostreig Estadístic (2n)
    - Disseny d'Experiments (2n)
    - Simulació, Remostreig i Aplicacions (3r)
    - Estadística Oficial i Disseny d'Enquestes (2n)
    - Sèries Temporals i Predicció (3r)
- **Clúster 4**:
    - Bases de Dades (2n)
    - Aplicacions de l'Estadística a la Bioinformàtica (3r)
    - Càlcul (1r)
    - Eines Informàtiques per a l'Estadística (1r)
    - Estadística Descriptiva (1r)
    - Mètodes Algebraics per a l'Estadística (1r)
    - Temes de Biociència i Ciència (1r)
    - Models Lineals (2n)
    - Inferència Estadística II (2n)
    - Optimització i Processos Estocàstics (2n)
    - Anàlisi de la Supervivència i Fiabilitat (2n)
    - Distribucions Multidimensionals (2n)

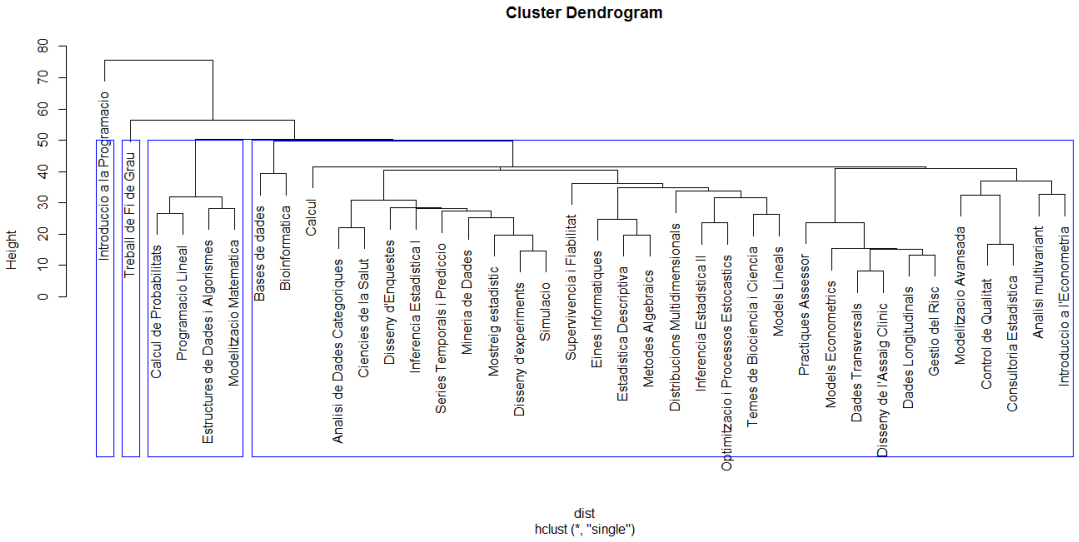
Veiem que amb el mètode complete, trobem uns clústers bastant interessants pel que fa a la homogeneïtat dins del clúster en relació al curs acadèmic.

4.3.2 Mètode Single

També conegut com mètode del *veï més proper*, en aquest mètode es considera que la distancia o similtud entre dos clústers ve donada, respectivament, per la mínima distància (o màxima similitud) entre els seus components.

.

FIGURA 17. DENDOGRAMA PEL MÈTODE SINGLE ASSIGNANT 4 CLUSTERS ETIQUETANT PER NOM D'ASSIGNATURA



Aquesta partició del dendograma no ens és gaire útil ja que la distancia entre clústers es més gran i la repartició dins dels clúster es heterogénea. El mètode single és molt sensible als valors outliers com podem observar en la formació dels dos primers clúster son agrupa només una asignatura.

4.3.3 Mètode Average

Aquí la distància entre 2 clústers vindrà donada per la mitjana dels seus components.

FIGURA 18. DENDOGRAMA PEL MÈTODE AVERAGE ASSIGNANT 4 CLUSTERS ETIQUETANT PER NOM D'ASSIGNATURA

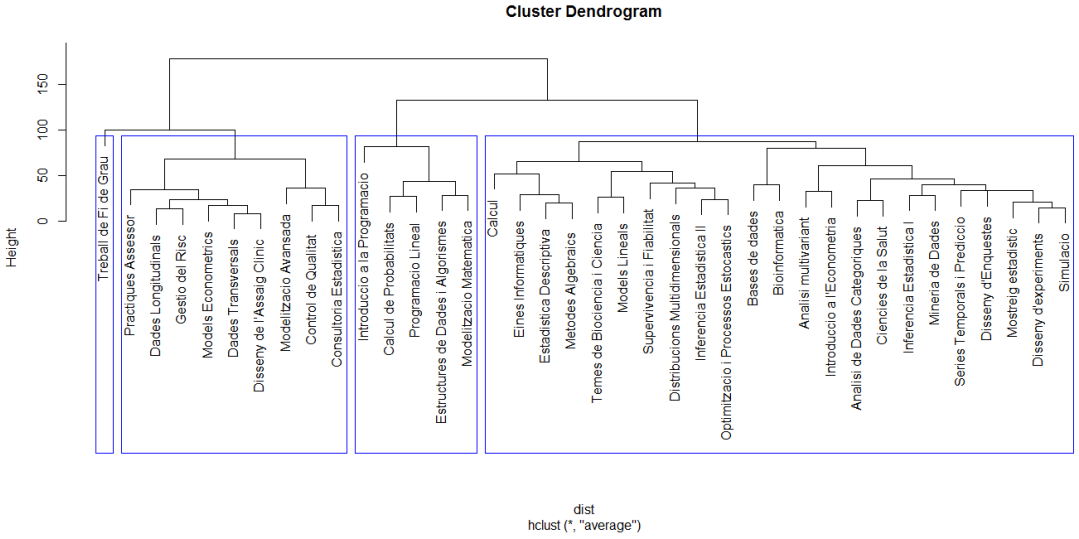
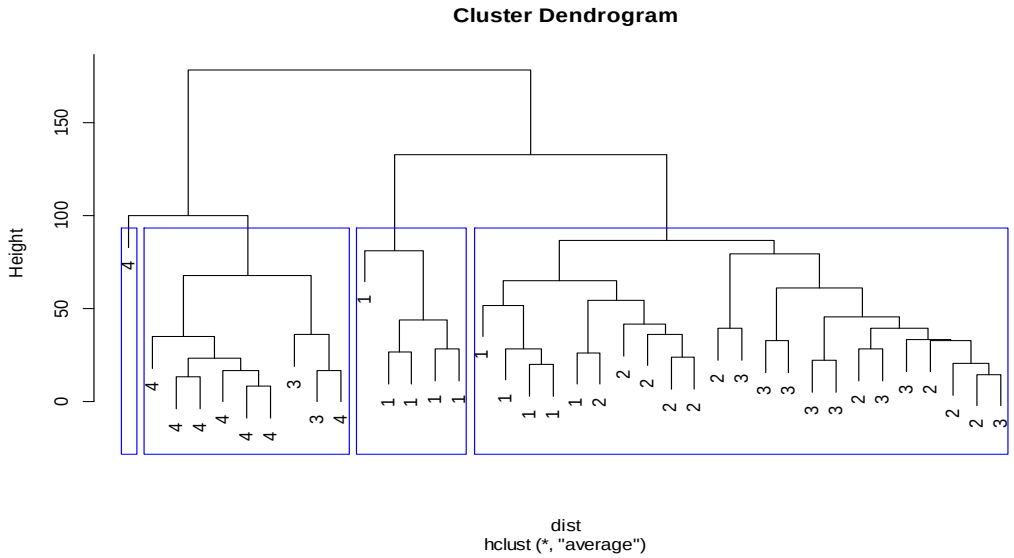


FIGURA 19. DENDOGRAMA PEL MÈTODE AVERAGE ASSIGNANT 4 CLUSTERS ETIQUETANT PER CURS



Amb el mètode average, torna a ser sensible a outliers ja que fer servir un clúster per a només una assignatura i això no ens serveix gaire de cara a una visualitació representativa de les agrupacions. Fa agrupacions millors que el mètode single però no ens serveix.

4.3.4 Mètode Ward

El mètode de Ward és un procediment jeràrquic en el qual, s'uneixen els dos clústers pels quals es tingui el menor increment en el valor total de la suma de quadrats de les diferències, dins de cada clúster, de cada individu al centroide del clúster.
La solució amb menor suma de quadrats total és l'escollida.

FIGURA 20. DENDOGRAMA PEL MÈTODE WARD ASSIGNANT 4 CLUSTERS ETIQUETANT PER NOM D'ASSIGNATURA

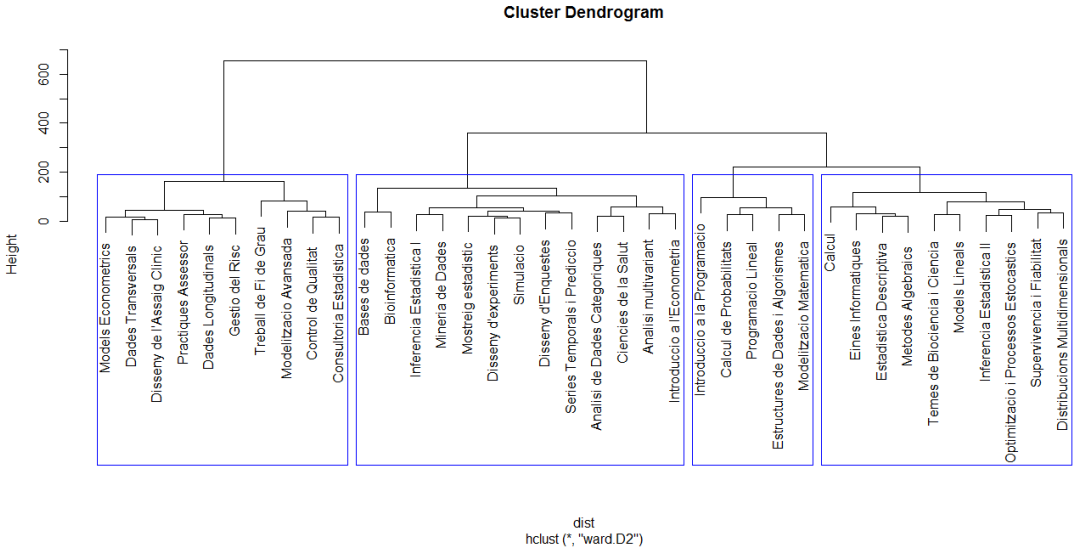
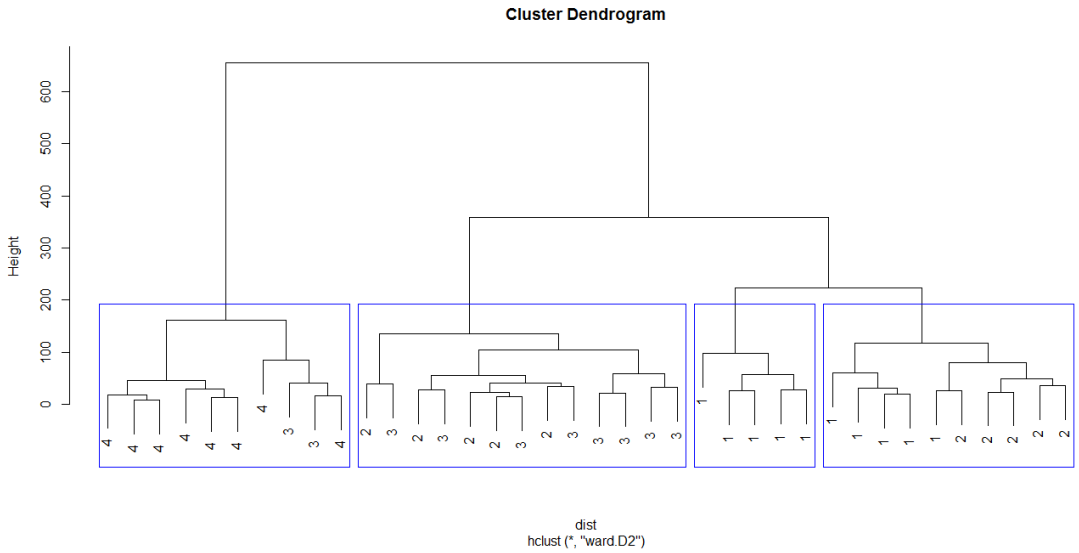


FIGURA 21. DENDOGRAMA PEL MÈTODE WARD ASSIGNANT 4 CLUSTERS ETIQUETANT PER CURS



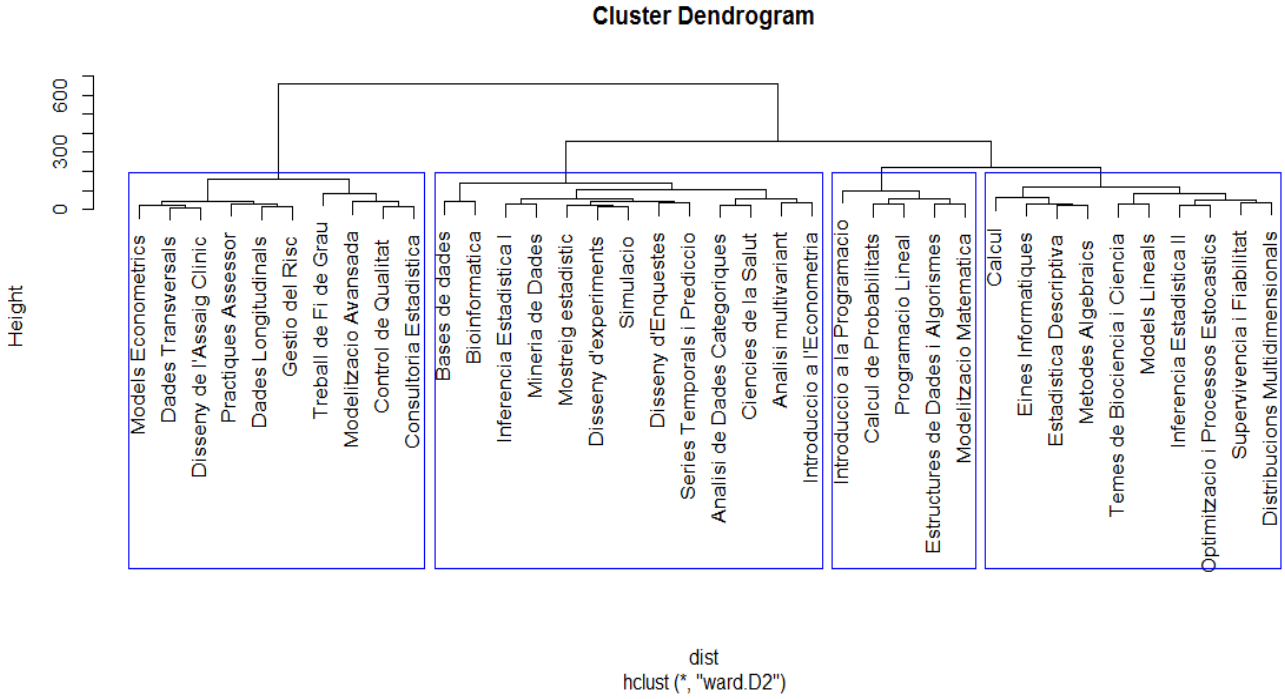
Aquest dendogroma, sembla força representatiu i té certa coherència, i ens fixem en les agrupacions, són quasi iguals que les del mètode *complete*, són les següents:

- **Clúster 1**:
    - Models Economètrics (4rt)
    - Dades Transversals: Temes Avançats en Ciències de la Salut (4rt)
    - Disseny de l'Assaig Clínic i Metodologia Estadística Aplicada (4rt)
    - Pràctiques Externes (4rt)
    - Dades Longitudinals: Temes Avançats en Ciències de la Salut (4rt)
    - Gestió del Risc (4rt)
    - Treball de Fi de Grau (4rt)
    - Modelització Avançada (3)
    - Control de Qualitat i Estadística Industrial (3)
    - Consultoria Estadística (4rt)
- **Clúster 2:**
    - Bases de dades (2n)
    - Aplicacions de l'Estadística a la Bioinformàtica (3r)
    - Inferència Estadística I (2n)
    - Mineria de dades (3r)
    - Mostreig estadístic (2n)
    - Disseny d'experiments (2n)
    - Simulació, Remostreig i Aplicacions (3r)
    - Estadística Oficial i Disseny d'Enquestes (2n)
    - Sèries Temporals i Predicció (3r)
    - Anàlisi de Dades Categòriques (3r)
    - Aplicacions de l'Estadística a les Ciències de la Salut (3r)
    - Anàlisi Multivariant (3r)
    - Introducció a l'Econometria (3r)
- **Clúster 3**:
    - Introducció a la Programació (1r)
    - Càlcul de Probabilitats (1r)
    - Programació Lineal (1r)
    - Estructures de Dades i Algorismes (1r)
    - Modelització Matemàtica (1r)
- **Clúster 4:**
    - Càlcul (1r)
    - Eines Informàtiques per a l'Estadística (1r)
    - Estadística Descriptiva (1r)
    - Mètodes Algebraics per a l'Estadística (1r)
    - Temes de Biociència i Ciència (1r)
    - Models Lineals (2n)
    - Inferència Estadística II (2n)
    - Optimització i Processos Estocàstics (2n)
    - Anàlisi de la Supervivència i Fiabilitat (2n)
    - Distribucions Multidimensionals (2n)

4.4 Elecció del dendrograma

Tenint en compte, les distàncies entre clústers i la coherència en les agrupacions relacionada amb els cursos acadèmics, els millor dendrogrames és el que hem obtingut amb el mètode de ***ward.***

FIGURA 22. DENDOGRAMA SELECCIONAT



Com hem comentat abans, realitza agrupacions força coherents entre elles, al primer agrupa 8 assignatures de 4rt i 2 de 3r, amb un comportament semblant en relació al rendiment acadèmic si observem els gràfics bivariants del capítol dos. Les agrupacions pertanyents al clúster 1 tenen rendiment alt.

Al segon clúster barreja assignatures de segon i tercer curs amb rendiment mig alt si les observem amb detall al capítol 2.

Al tercer clúster agrupa 5 assignatures amb rendiment per sota del 50%. Les hem observat amb deteniment al capítol dos, té coherència que s'ajustin en un clúster ja que són molt homogènies entre elles i molt heterogènes amb les demés.

Al quart clúster agrupa assignatures de 1r i 2n on com ja sabem el rendiment es mig. Per tant torna a haver-hi homogeneïtat dins del clúster i heterogeneïtat entre els clústers. Que és el que buscàvem a primera instància.

4.5 Algorisme K-Means

L'anàlisi de conglomerats de *k-means*, igual que altres tècniques d'anàlisis de grups, tracta d'identificar grups de casos homogenis (individus o objectes) que tinguin comportaments, característiques o atributs similars.

L'objectiu de l'algorisme és obtenir k grups de manera que es minimitzi la suma de quadrats intra-grups (suma dels quadrats de les diferències entre els valors de les variables observades en cada individu de la mostra respecte dels valors mitjos del grup al que pertany).

Cada individu és assignat inicialment al grup amb el qual presenta menor distància (mesura per la distància euclidiana als valors mitjos del grup). L'assignació de cada individu es modifica i se l´integra en un altre grup si amb això s'aconsegueix una reducció en la suma de quadrats intragrups.

Cada vegada que es reassigna un individu a un altre grup es calculen de nou les mitjanes del grup en totes les variables seleccionades. Les reassignacions finalitzen quan ja no es produeix cap transferència entre grups o s'ha arribat a el nombre màxim d'iteracions permeses

Procedim a realitzar l'algorisme k-means per a les nostres dades amb R amb el següent resultat:

Mida dels clústers:

K-means clustering with 4 clusters of sizes 12, 6, 12, 8

Veiem com s'ha realitzat dos clústers de mida 12. És a dir, dos grups de 12 assignatures, i dos clústers més petits , un de 6 assignatures i un de 8.

Mitjanes de les variables dins del clúster

Cluster means:

TAULA 6. CENTROIDES DELS CLUSTERS

Cluster	Matriculats	mh	Exc	Nt	Ap	S	Np	Rendi	Superats	Presentats
1	169.50	3.58	5.25	36.83	60.91	24.25	38.66	63.32	106.58	130.83
2	236.83	5.16	4.83	31.00	65.66	35.66	94.50	45.37	106.66	142.33
3	120.25	3.83	10.41	28.00	40.33	13.75	23.91	68.89	82.58	96.33
4	42.87	1.62	8.12	16.37	11.50	1.00	4.25	86.76	37.62	38.62

Aquesta taula és força interessant ja que observem com de semblants o diferents són les mitjanes de les variables de cada clúster.

Per sortir de dubtes, realitzarem un ANOVA per veure si ha diferències entre el rendiment per clústers.

TukeyHSD(ANOVA1)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = rendi ~ clusters)

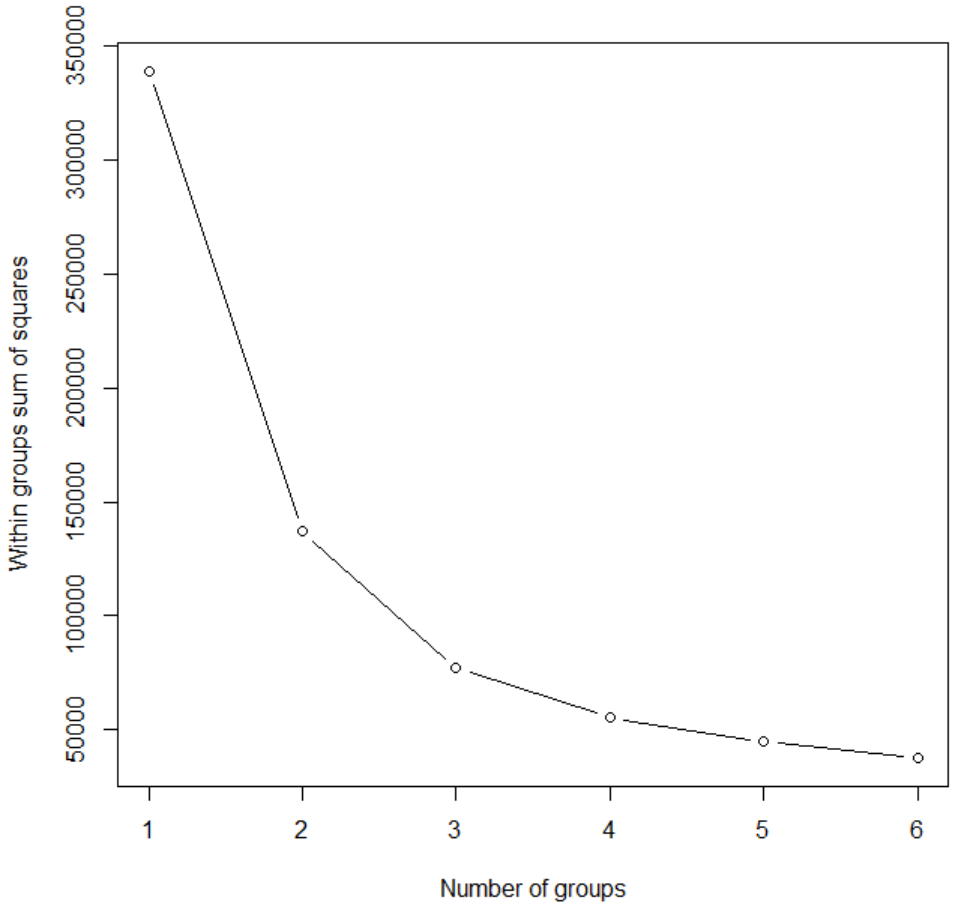
$clusters

	diff	lwr	upr	p adj
2-1	-16.04515	-26.54198	-5.548312	**0.0012261**
3-1	-26.57774	-37.45795	-15.697533	**0.0338787**
4-1	-41.85346	-55.35545	-28.351476	**0.0000009**
3-2	-10.53260	-20.44953	-0.615662	**0.0000241**
4-2	-25.80832	-38.54685	-13.069781	**0.0000000**
4-3	-15.27572	-28.33197	-2.219466	**0.0166211**

Observem com per cada comparació de clústers, el p-valor és significatiu, els que ens porta a tenir prou evidència de rebutjar l' hipòtesis nul.la i afirmar que les mitjanes del rendiment entre clústers són diferents.

Fem un gràfic de les variabilitats *whithin* si féssim ara entre 2 i 6 conglomerats amb el mètode k-*means* per veure amb quants grups ens quedaríem.

FIGURA 23. GRÀFIC DEL NOMBRE DE GRUPS SEGONS LA VARIABILITAT WITHIN



El gràfic ens indica que k=4 és la quantitat adequada de clústers, com havíem valorat a l'inici.

5. Arbres de Regressió

Una alternativa al model de regressió lineal múltiple pot ser l'arbre de regressió.

Els arbres de regressió són una tècnica estadística amb la qual volem cercar un bon model. En el nostre cas un bon model explicatiu.

Volem modelar la variable resposta *Rendiment* a partir d'un conjunt de variables explicatives. Aquest mètode ens aporta uns resultats fàcils de interpretar i considera l'efecte de possibles iteracions entre les variables explicatives de forma natural.

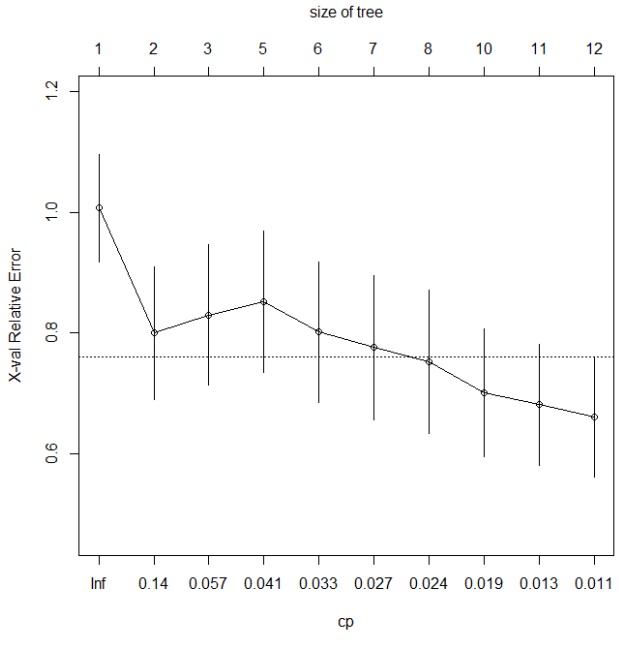
Aquesta tècnica consisteix, a grans trets, en dividir l'espai de les variables independents de forma que els valors de la variable resposta siguin cada cop mes homogenis dintre de cadascuna de les diferents particions, i que per tant minimitzin la suma de residus al quadrat.

Gràcies a la seva representació gràfica els arbres permeten una visió molt clara i explicativa del model generat, fet que valorarem durant l'estudi amb les nostres dades.

L'algoritme utilitzat per a produir l'arbre de regressió permet dividir l'arbre en tantes fulles com individus tingui la mostra, de forma que cada individu seria representat per una fulla concreta amb unes característiques molt determinades. Per tal de no fer un us excessiu d'aquest algoritme, degut al temps computacional que aquest comporta i a la sobre-parametrització de l'arbre s'han estipulats diferents criteris de parada.

El ***primer criteri*** que s'ha utilitzat es coneix com l'obtenció del paràmetre de complexitat. Aquest criteri mostra de forma gràfica el valor de l'error relatiu en funció del nombre de particions, i ens ajuda a escollir el llindar del mínim error *cross-validat*. Per a l'aplicació d'aquest criteri ens hem ajudat del següent gràfic en el que es podrà visualitzar el mencionat llindar.

FIGURA 24. GRÀFIC DEL LLINDAR MÍNIM CROSS-VALIDAT



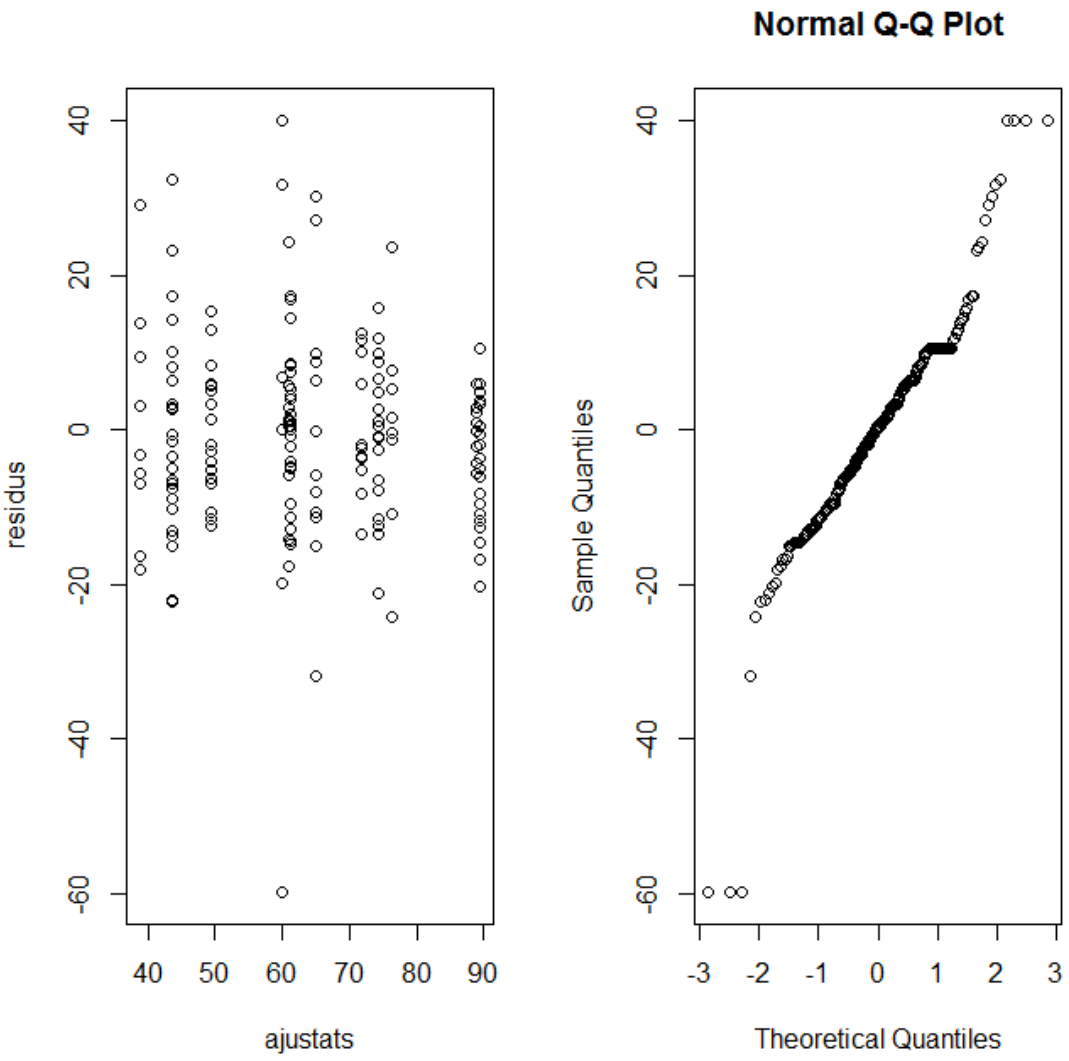
Tenim dues opcions, en primer lloc, podem escollir aquell que tingui l'error relatiu mínim, en el nostre cas 0.011. Un segon criteri consisteix en fixar un llindar en el mínim dels errors cross-validat més una desviació estàndard, i aleshores escollir el primer arbre amb un error inferior al llindar, en el nostre cas 0.01).

El ***segon criteri*** basat en les validacions creuades que realitza l'*R*, l'arbre òptim que escolliríem correspondria a escollir on s'ajunten definitivament l'$R2intern$ i $R2extern$

5.1 Avaluació dels residus:

- Avaluació de la normalitat i homocedasticitat dels residus.

FIGURA 25. AVALUACIÓ DEL COMPORTAMENT DELS RESIDUS



No observem incongruències.

5.2 Arbre seleccionat

FIGURA 25. SORTIDA D'R DE L'ARBRE DE REGRESIÓ SELECCIONAT

```
Regression tree:
rpart(formula = rendi ~ ., data = dades, cp = 0.011)

Variables actually used in tree construction:
[1] ap          Curs        Matriculats nt          s

Root node error: 109644/228 = 480.89

n= 228

         CP nsplit rel error  xerror     xstd
1  0.283179      0   1.00000 1.01553 0.089621
2  0.072386      1   0.71682 0.78866 0.115934
3  0.044280      2   0.64444 0.75507 0.118632
4  0.037135      4   0.55588 0.76218 0.121644
5  0.029083      5   0.51874 0.74754 0.121770
6  0.024769      6   0.48966 0.70868 0.116163
7  0.024090      7   0.46489 0.70331 0.115401
8  0.014438      9   0.41671 0.66878 0.117708
9  0.012255     10   0.40227 0.64868 0.116678
10 0.011000     11   0.39002 0.64845 0.116693
```

Com veiem a la sortida d'R, tenim 10 particions amb 12 nodes.

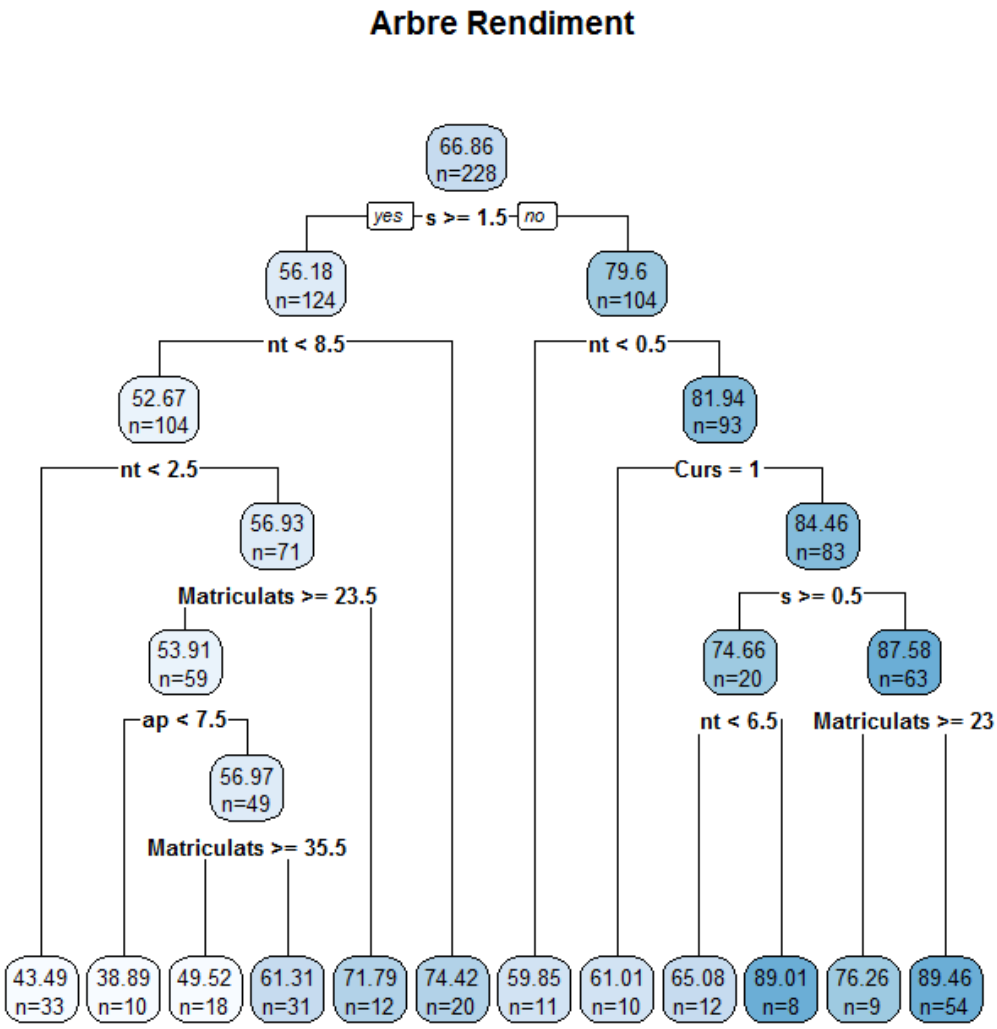
Ens interessen els valors mínim de rel error i xerror per trobar els R quadrats intern:

- R2RT<-SSR/SST
- R2RT
- [1] 0.6099836

En efecte 0.6099836 és el resultat de restar 1 al mínim error relatiu:

$$1 - 0.39002 = \mathbf{0.6099836}$$

FIGURA 26. ARBRE DE REGRESIÓ SELECCIONAT



Observem que hem obtingut un arbre amb un total de 12 nodes generats a partir de 10 particions.

El nombre de ***suspesos*** és la variable que millor discrimina el rendiment, tot i que dintre de la mateixa branca podem apreciar altres diferències.

L'arbre també recurreix molt a la variable ***Matriculats*** per discriminar entre rendiments, amb el que s'aprecia com a menys matriculats, millor rendiment

El rendiment més baix de l'arbre s'observa quan el nombre de suspesos és igual o superior a 1.5, el nombre de matriculats és més gran de 23.5 i el nombre d'aprovats és més petit que 7.5.

Mentre que el rendiment més alt, és en els cursos 2, 3 o 4, amb 0 suspesos i menys de 23 matriculats.

Amb lo qual podem veure com a menys matriculats millor rendiment.

5.3 Validació de l'arbre de regressió

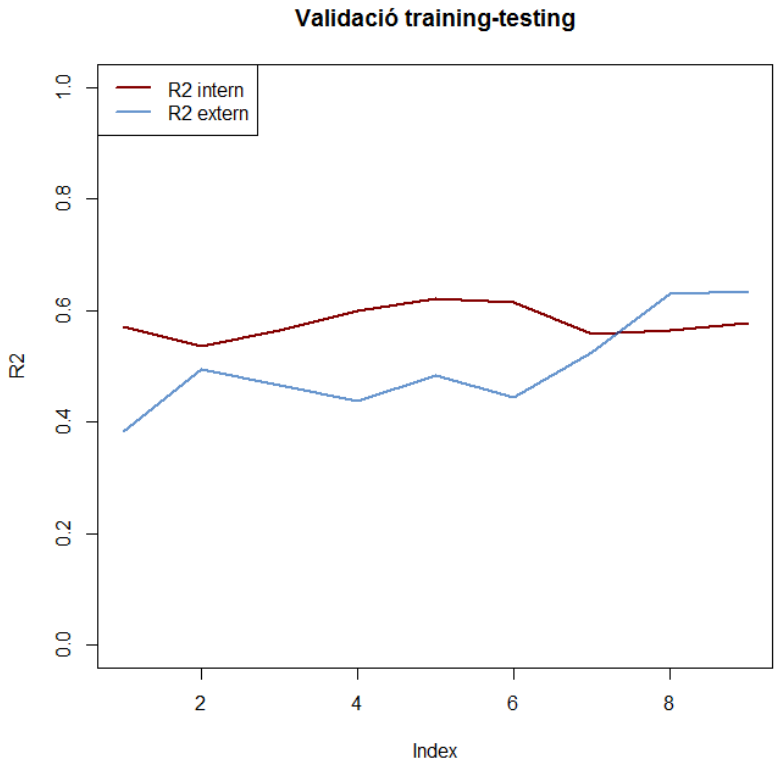
5.3.1 Validacio mida testing training

Per validar l'arbre de regressió que hem obtingut utilitzarem el mètode de training i testing. Aquest mètode, consisteix en dividir la mostra inicial en dues parts, la primera anomenada "training" la utilitzarem per generar el nostre arbre de regressió, amb aquesta podrem avaluar l'error intern. La segona part de la mostra, anomenada "testing" s'utilitza per avaluar si l'arbre escollit realment prediu adequadament per a unes altres dades (error extern), ja que, les dades testing no han estat utilitzades a l'hora de crear l'arbre.

La mida habitual d'aquestes de la mostra training es del 75%, i per a la mostra testing del 25%. A continuació hem comprovat si realment la mida d'aquestes dues mostres afecta als resultat. Per això, hem simulat l'error tant intern com extern (coeficient de determinació) amb una mida mostral per al training del 10% fins al 90%.

Al següent, podem veure que el percentatge de mostra training-testing afecta relativament alhora de predir. Veiem que en quasi tot moment el coeficient de determinació intern és més elevat que l'extern, tal i com era d'esperar. El nostre arbre explica de mitjana un 0. 6119647 internament i un 0. 4840028 de forma externa.

FIGURA 27. VALIDACIÓ TRAINING TEST



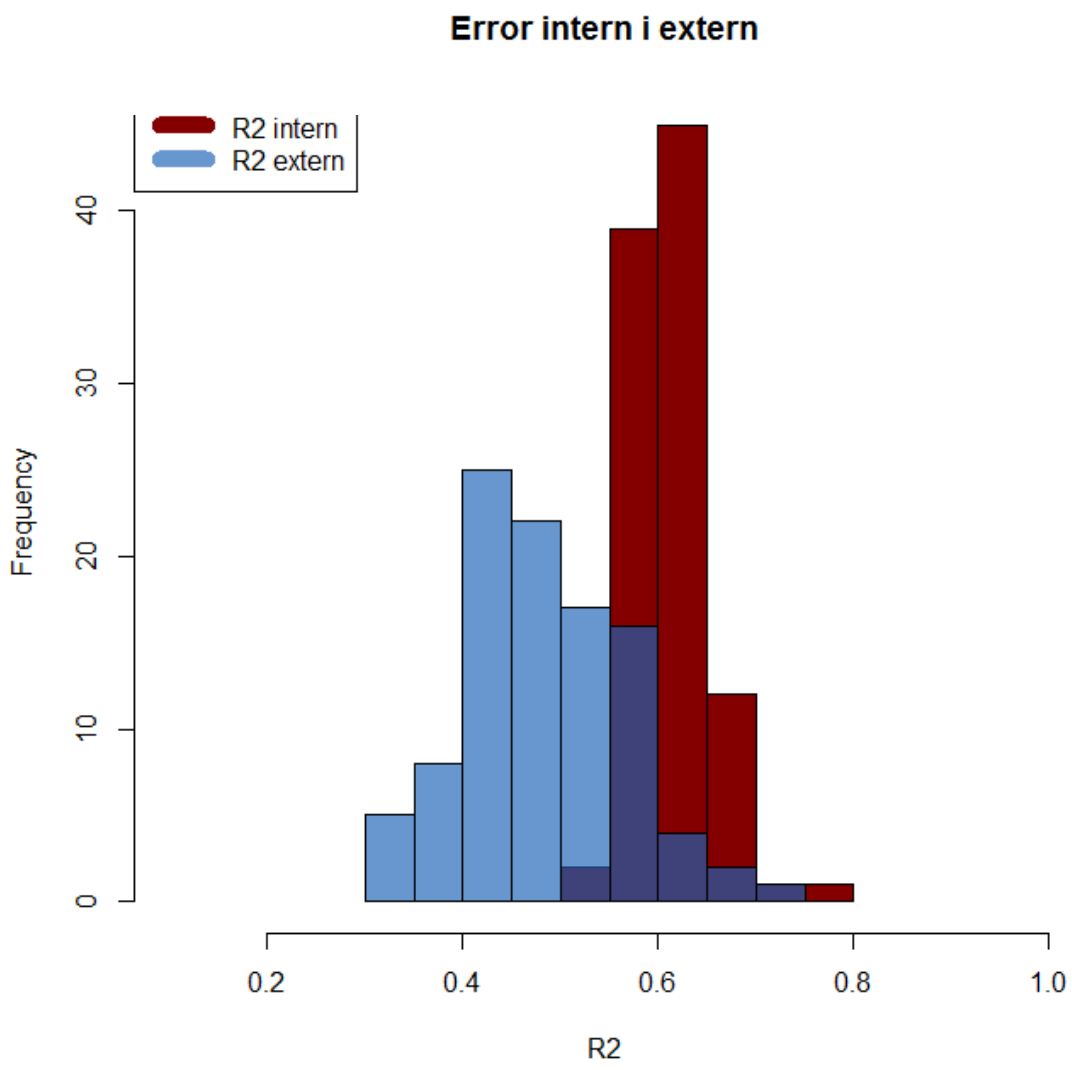
5.3.2 Validació llavor mostra

També hem volgut analitzar si la llavor de la mostra per escollir el training-testing afecta als resultat. Per això, hem fixat el valor que s'acostuma a utilitzar, ja que, hem vist que no hi ha massa diferència. Tot i així, veiem que agafant només un 50% de mostra training obtindríem els mateixos resultats.

Al gràfic es mostren les freqüències per als diferents coeficients de determinació. Corroborem que l'error extern sempre és menor que l'intern. I que depenent de la llavor, podem obtenir per a l'error intern un mínim de 0.55 i un màxim de 0.69. L'error extern varia de 0.33 a 0.65.

Aquest gràfic mostra que el model s' ajusta prou bé, és a dir, són arbres útils per predir rendiment acadèmic.

FIGURA 28. VALIDACIÓ LLAVOR MOSTRA. R2 INTERN I R2 EXTERN



6. Conlusions

Les principals conclusions a les que s'ha arribat realitzant aquest estudi són:

- Les assignatures amb rendiment més baix són:

    - Càlcul de Probabilitats.
    - Modelització Matemàtica.
    - Estructura de Dades i Algorismes.
    - Introducció a la Programació.
    - Anàlisi de la Supervivència i Fiabilitat.

- El 80% de les assignatures amb rendiment menor al 50%, és a dir, 4 de 5, es troben a primer curs.

- No hi ha diferències estadísticament significatives entre el rendiment del 1r semestre i el del 2n.

- El rendiment augmenta a mesura que augmenta el curs acadèmic.

- No hi ha tendència anual en el rendiment acadèmic.

- Realitzant anàlisi de clústers per veure quines assignatures tenen un comportament més semblant hem escollit les agrupacions del mètode ward amb els següents 4 clústers:

    - **Clúster 1**:
        - Models Economètrics (4rt)
        - Dades Transversals: Temes Avançats en Ciències de la Salut (4rt)
        - Disseny de l'Assaig Clínic i Metodologia Estadística Aplicada (4rt)
        - Pràctiques Externes (4rt)
        - Dades Longitudinals: Temes Avançats en Ciències de la Salut (4rt)
        - Gestió del Risc (4rt)
        - Treball de Fi de Grau (4rt)
        - Modelització Avançada (3)
        - Control de Qualitat i Estadística Industrial (3)
        - Consultoria Estadística (4rt)
    - **Clúster 2:**
        - Bases de dades (2n)
        - Aplicacions de l'Estadística a la Bioinformàtica (3r)
        - Inferència Estadística I (2n)
        - Mineria de dades (3r)
        - Mostreig estadístic (2n)

        - Disseny d’experiments (2n)
        - Simulació, Remostreig i Aplicacions (3r)
        - Estadística Oficial i Disseny d'Enquestes (2n)
        - Sèries Temporals i Predicció (3r)
        - Anàlisi de Dades Categòriques (3r)
        - Aplicacions de l'Estadística a les Ciències de la Salut (3r)
        - Anàlisi Multivariant (3r)
        - Introducció a l'Econometria (3r)
    - **Clúster 3**:
        - Introducció a la Programació (1r)
        - Càlcul de Probabilitats (1r)
        - Programació Lineal (1r)
        - Estructures de Dades i Algorismes (1r)
        - Modelització Matemàtica (1r)
    - **Clúster 4:**
        - Càlcul (1r)
        - Eines Informàtiques per a l'Estadística (1r)
        - Estadística Descriptiva (1r)
        - Mètodes Algebraics per a l'Estadística (1r)
        - Temes de Biociència i Ciència (1r)
        - Models Lineals (2n)
        - Inferència Estadística II (2n)
        - Optimització i Processos Estocàstics (2n)
        - Anàlisi de la Supervivència i Fiabilitat (2n)
        - Distribucions Multidimensionals (2n)

- Un cop realitzat l’arbre de regresió hem vist com la variable que millor discrimina el rendiment és el nombre de suspesos junt amb la del nombre de matriculats.

- A menys nombre de Matriculats, més rendiment.

7. Bibliografia

- Everitt, B., Hothorn, T. : An introduction to Applied Multivariate Analysis with R. Springer. 2011.

- Peña, D.: Análisis de datos multivariantes. McGraw Hill. 2002.

- Rencher, A., Christensen, W.: Methods of Multivariate Analysis. Wiley Series in Probability and Mathematical Statistics. 2012.

- Links:

    - https://www.r-bloggers.com/performing-anova-test-in-r-results-and-interpretation/

    - https://rpubs.com/minma/cart_with_rpart

8. Annex

```
########## Codi R #################

install.packages("lmtest")
install.packages("xlsx")
install.packages("tseries")
install.packages("openxlsx")
install.packages("ggplot2")
install.packages("rJava")
install.packages("tree")
install.packages("rpart")
install.packages("rpart.plot")
library(tree)
library(lmtest)
library(rJava)
library(xlsx)
require(xlsxjars)
library(tseries)
library(openxlsx)
library(ggplot2)
library(rpart)
library(rpart.plot)

############################
#BIVARIANT#

dades<-read.xlsx("G:/dades/dadesTFG.xlsx")
attach(dades)

#Transformem variables les variables categoriques a factors#

dades$Curs<-as.factor(dades$Curs)
dades$Semestre<-as.factor(dades$Semestre)
dades$Any<-as.factor(dades$Any)
dades$Codi<-as.factor(dades$Codi)
dades$Assignatura<-as.factor(dades$Assignatura)

#Explorem relació variables#

plot(rend ~ Curs,ylab="Rendiment",main="Rendiment per Curs", dades)##

plot(rend~ Any,ylab="Rendiment",main="Rendiment per Any", dades)

dades$Any<-as.numeric(dades$Any)#Per fer el model
fit <- lm(rend ~ Any, data = dades)
summary(fit)

plot(rend~ Semestre,ylab="Rendiment",main="Rendiment per Semestre", dades)
```

```
#Fem un t.test per assegurar-nos##
t.test(rend~ Semestre, data=dades)

#####################################################################

####CLUSTERS####

#llegim dades##

dades<-read.xlsx("G:/dades/dadescluster.xlsx")

attach(dades)

Curs<-as.factor(Curs)
Semestre<-as.factor(Semestre)
Assignatura<-as.factor(Assignatura)

dist<-dist(dades,method="euclidean")#Calculem distancia euclidiana

## Complete ##

csc <- hclust(dist, method="complete")
plot(csc,labels=Assignatura,main="Complete") # mostra dendogram
groups <- cutree(csc, k=4) # separem en 4 clusters

# Marquem amb un quadre blau les formacions que ens assigna

rect.hclust(csc, k=4, border="blue")
##Complete label curs##
csc <- hclust(dist, method="complete")
plot(csc,labels=Curs) # mostra dendogram
groups <- cutree(csc, k=4) # separem en 4 clusters

# Marquem amb un quadre blau les formacions que ens assigna

rect.hclust(csc, k=4, border="blue")

## Single ##

css <- hclust(dist, method="single")
plot(css,labels=Assignatura) # mostra dendogram
groups <- cutree(css, k=4) # separem en 4 clusters
# Marquem amb un quadre blau les formacions que ens assigna

rect.hclust(css, k=4, border="blue")

## Average ##

csa <- hclust(dist, method="average")
plot(csa,labels=Assignatura) # mostra dendogram
groups <- cutree(csa, k=4) # separem en 4 clusters
#Marquem amb un quadre blau les formacions que ens assigna
rect.hclust(csa, k=4, border="blue")

## Average labelcurs ##
```

```
csa <- hclust(dist, method="average")
plot(csa,labels=Curs) # mostra dendogram
groups <- cutree(csa, k=4) # separem en 4 clusters
# Marquem amb un quadre blau les formacions que ens assigna
rect.hclust(csa, k=4, border="blue")

## Ward ## és el millor mètode dels 4
csw <- hclust(dist, method="ward.D2")
plot(csw,labels=Assignatura) # mostra dendogram
groups <- cutree(csw, k=4) # separem en 4 clusters
rect.hclust(csw, k=4, border="blue")

## Ward label curs ## és el millor mètode dels 4

csw <- hclust(dist, method="ward.D2")
plot(csw,labels=Curs) # mostra dendogram
groups <- cutree(csw, k=4) # separem en 4 clusters
rect.hclust(csw, k=4, border="blue")


###KMEANSS####

##Teiem les variables Assignatura, Semestre i ANy, per treballar amb només dades continues

dadesk<-dades[,-c(1,12:13)]

##Fem els 4 k-means (el que canvia és l'algoritme)##

kmeans(dadesk, 4, iter.max = 10, nstart = 1,algorithm = "Hartigan-Wong")###83.7
kmeans(dadesk, 4, iter.max = 10, nstart = 1,algorithm = "Lloyd")##83.7
kmeans(dadesk, 4, iter.max = 10, nstart = 1,algorithm = "Forgy")##80.9
kmeans(dadesk, 4, iter.max = 10, nstart = 1,algorithm = "MacQueen")##80.6


clusters <-kmeans(dadesk, 4, iter.max = 10, nstart = 1,algorithm = "Hartigan-Wong")$centers

#centroides del 4 grups

clusters

#Comprovar que el rendiment es diferent en cada cluster

ANOVA1 <- aov(rendi ~ clusters, data=dadesk)

TukeyHSD(fit)

##Busca la variabilitat whithin dels conglomerats que hem format##

##Si fessim fins a 6 grups, a partir de quants en tindriem prou#

wss <- c();
for (i in 1:6)
{wss[i] <- sum(kmeans(dadesk, centers = i)$withinss)}
plot(1:6, wss, type = "b", xlab = "Number of groups",ylab = "Within groups sum of squares")
```

```
##Amb 4 grups tenim prou

################################################################################

###################
#ARBRES de Regresió#
###################

dades<-read.xlsx("G:/dades/dadesTFG.xlsx")

attach(dades)

#Treiem variables que no ens interesen

dades$Codi <- NULL
dades$rend <- NULL
dades$Assignatura <- NULL
dades$exit <- NULL
dades$fracas <- NULL
dades$Superats <- NULL
dades$Presentats <- NULL
dades$np<- NULL

#Transformem variables les variables categoriques a factors#
dades$Curs<-as.factor(dades$Curs)
dades$Semestre<-as.factor(dades$Semestre)
dades$Any<-as.factor(dades$Any)
arbre<-rpart(rendi~.,data=dades)
plotcp(arbre)#0.011


############################

#Avaluacio dels residus ##

data.frame(obs=rendi,pred=predict(arbre),resid=resid(arbre))
par(mfrow=c(1,2))
plot(predict(arbre),resid(arbre),xlab="ajustats",ylab="residus")
qqnorm(resid(arbre))

# Criteris#

taula.rend <- arbre$cptable

taula.rend <- data.frame(taula.rend)

nfulles <- taula.rend$nsplit+1

R2 <- 1-taula.rend$xerror

R2cross <- 1-taula.rend$xerror

CritAt <- taula.rend$xerror+taula.rend$xstd
```

```
# Criteri 1: Ordenem la columna xerror de la taula i treiem el primer valor que correspon al
#mínim.

OptimCriteri1 <- taula.rend$CP[order(taula.rend$xerror)[1]];OptimCriteri1

# Criteri 2:

OptimCriteri2 <- taula.rend$CP[order(taula.rend$xerror,decreasing=T)[1] <
order(CritAt)[1]];OptimCriteri2

taula.rend$fulles <- nfulles
taula.rend$R2 <- R2
taula.rend$R2cross <- R2cross
taula.rend$Criteri1 <- OptimCriteri1
taula.rend$Criteri2 <- OptimCriteri2

P<-predict(arbre)
SSE<-sum((rendi-P)^2)
SSR<-sum((P-mean(rendi))^2)
SST<-SSR+SSE
R2RT<-SSR/SST
##############
#validacio creuada
printcp(rpart(rendi~.,data=dades,cp=0.011))
arbre.prune <- prune.rpart(arbre,0.01)
rpart.plot(arbre.prune, main ="Arbre Rendiment",extra=1, type=2, digits=4)

### Validacio mida testing training ##
nSim <- 100
valI <- NULL
valE <- NULL
for(i in seq(0.1,0.9,by=0.01)) {

  set.seed(1)

  training <- sample(1:228,size=228*i)
  test <- (1:228)[! (1:228 %in% training)]
  dat.training <- dades[training,]
  dat.test <- dades[test,]
  dat.rpart <- rpart(rendi ~ .,data=dat.training,cp=0.01)

  P<-predict(dat.rpart)
  ptest<- predict(dat.rpart, newdata=dat.test)

  SSEi<-sum((dat.training$rendi-P)^2)
  SSRi<-sum((P-mean(dat.training$rendi))^2)
  SSTi<-SSRi+SSEi

  SSEe<-sum((dat.test$rendi-ptest)^2)
  SSRe<-sum((ptest-mean(dat.test$rendi))^2)
  SSTe<-SSRe+SSEe
```

```
  valI[i*10] <- 1-SSEi/SSTi
  valE[i*10] <- 1-SSEe/SSTe

}

mean(valE)
mean(valI)

plot(valI,type="l",col="#840000",ylim=c(0,1),lwd=2,ylab="R2",main="Validació training-
testing")
lines(valE,type="l",col="#1c61b6AA",lwd=2)
legend("topleft", c("R2 intern","R2 extern"),col=c("#840000","#1c61b6AA"),lty=1,lwd=2)

# Validacio llavor mostra###

set.seed(1)
nSim <- 100
valI <- NULL
valE <- NULL
for(i in 1:nSim) {

  training <- sample(1:228,size=228*(3/4))
  test <- (1:228)[! (1:228 %in% training)]
  dat.training <- dades[training,]
  dat.test <- dades[test,]
  dat.rpart <- rpart(rendi ~ .,data=dat.training)

  P<-predict(dat.rpart)
  ptest<- predict(dat.rpart, newdata=dat.test)

  SSEi<-sum((dat.training$rendi-P)^2)
  SSRi<-sum((P-mean(dat.training$rendi))^2)
  SSTi<-SSRi+SSE

  SSEe<-sum((dat.test$rendi-ptest)^2)
  SSRe<-sum((ptest-mean(dat.test$rendi))^2)

  SSTe<-SSRe+SSEe

  valI[i] <- 1-SSEi/SSTi

  valE[i] <- 1-SSEe/SSTe

}
R2int.RT<-paste(   round(mean(valI),2), "(" ,  round(quantile(valI,probs=c(0.025)),2), "-
",round(quantile(valI,probs=c(0.975)),2)  ,")")

R2ext.RT<-  paste(   round(mean(valE),2), "(" ,  round(quantile(valE,probs=c(0.025)),2), "-
",round(quantile(valE,probs=c(0.975)),2)  ,")")

par(mfrow=c(1,1))
hist(valI,main="Error intern i extern",xlim=c(0.10,0.99),col="#840000", xlab="R2")
hist(valE,add=T,main="Error extern",xlim=c(0.10,0.9),col="#1c61b6AA",ylim=c(0,30))
legend("topleft", c("R2 intern","R2 extern"),col=c("#840000","#1c61b6AA"),lty=1,lwd=10)
```